Adaptation-to-Context in Humans and Large Language Models

Inferring the underlying mechanisms of In-Context Learning with Structural Priming

Zhenghao Herbert Zhou, Yale University
Computational Cognitive Science Group, University of Maryland
May 5, 2026

Overview

Humans and LLMs are bidirectionally informative

Overview

Humans and LLMs are bidirectionally informative

Psycholinguistics

Structural Priming

Overview

Humans and LLMs are bidirectionally informative

Psycholinguistics

Computational Ling / NLP

Structural Priming

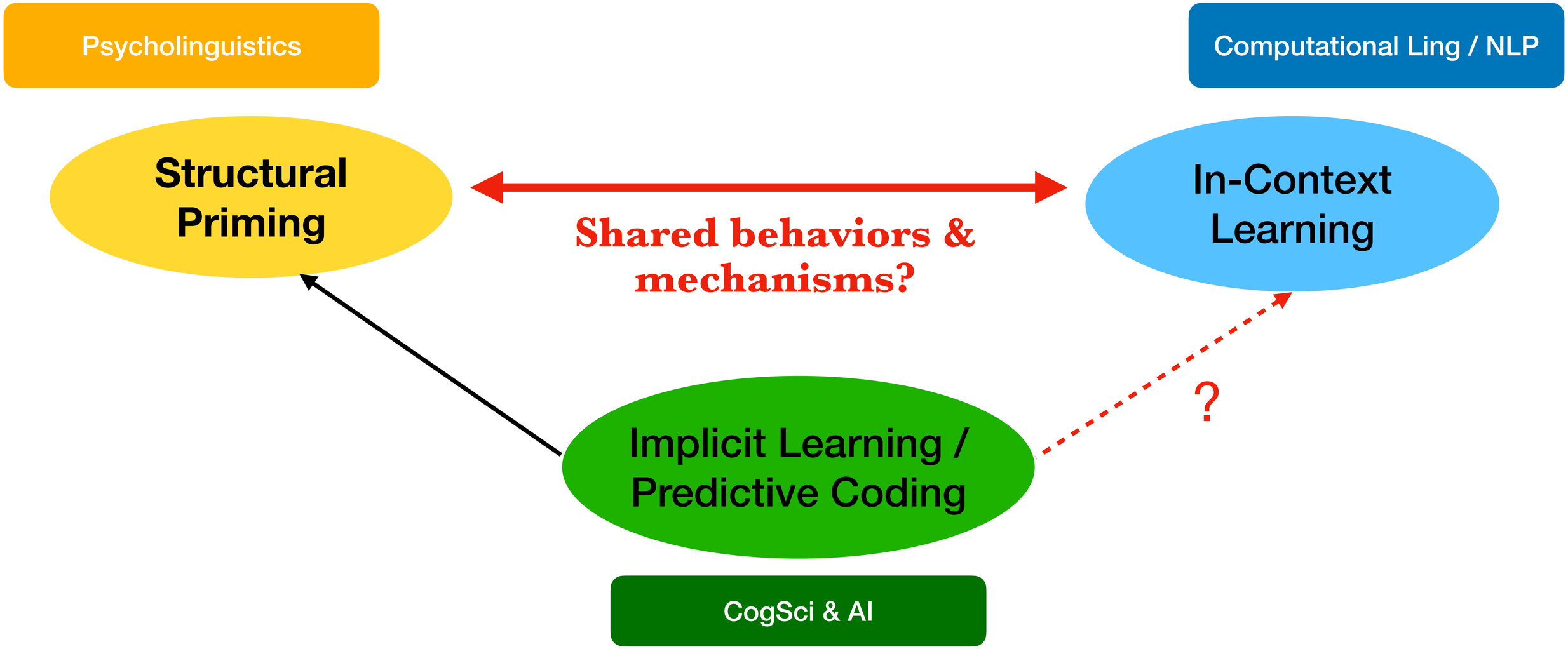
In-Context Learning

Overview

Humans and LLMs are bidirectionally informative

Overview

Humans and LLMs are bidirectionally informative



Structural Priming in Humans

Structural Priming

[e.g., Bock 1986; Chang 2012; etc.]

Structural Priming

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

[e.g., Bock 1986; Chang 2012; etc.]

Structural Priming

- **Structural Priming**: speakers tend to reuse the syntactic structures they have recently encountered during production or comprehension.
- A classical case: **dative alternation**
  - Double Object (DO): *Alice sent Bob a letter.*
  - Prepositional Dative (PD): *Alice sent a letter to Bob.*

[e.g., Bock 1986; Chang 2012; etc.]

Structural Priming

- **Structural Priming**: speakers tend to reuse the syntactic structures they have recently encountered during production or comprehension.
- A classical case: **dative alternation**
  - Double Object (DO): *Alice sent Bob a letter.*
  - Prepositional Dative (PD): *Alice sent a letter to Bob.*
- Other attested structures subject to priming:
  - active-passive;
  - *that*-omission in relative clauses and complement, etc.

[e.g., Bock 1986; Chang 2012; etc.]

The Inverse Frequency Effect (IFE)

- **Inverse Frequency Effect**: the *less preferred* (lower frequency) syntactic structure causes a *stronger* priming effect than the more preferred (higher frequency) structural alternative.

[e.g., Ferreira 2003, Jaeger & Snider 2007, 2013]

The Inverse Frequency Effect (IFE)

- **Inverse Frequency Effect**: the *less preferred* (lower frequency) syntactic structure causes a *stronger* priming effect than the more preferred (higher frequency) structural alternative.

[e.g., Ferreira 2003, Jaeger & Snider 2007, 2013]

The Inverse Frequency Effect (IFE)

- **Inverse Frequency Effect**: the *less preferred* (lower frequency) syntactic structure causes a *stronger* priming effect than the more preferred (higher frequency) structural alternative.

Verb Bias:

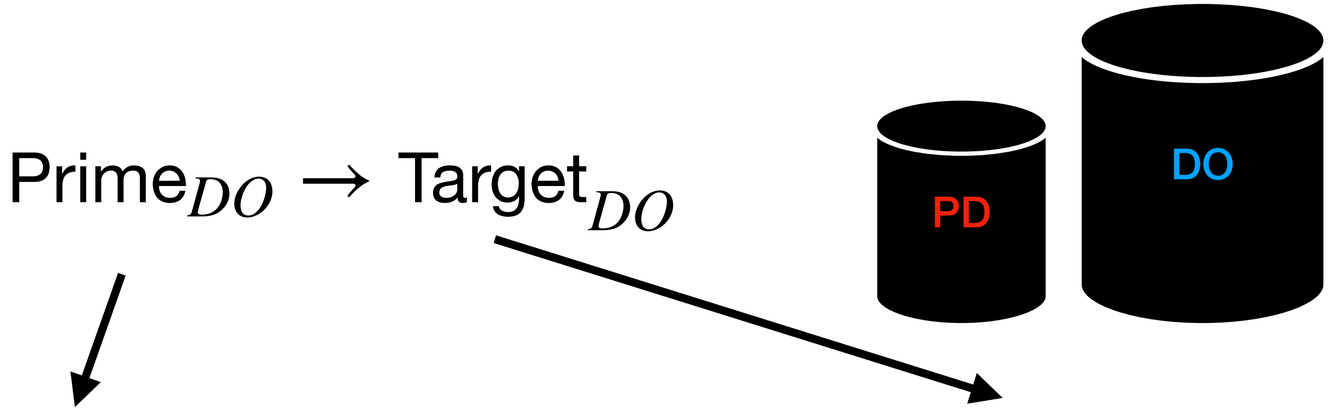
buy is biased towards DO

design is towards PD

[e.g., Ferreira 2003, Jaeger & Snider 2007, 2013]

The Inverse Frequency Effect (IFE)

- **Inverse Frequency Effect**: the *less preferred* (lower frequency) syntactic structure causes a *stronger* priming effect than the more preferred (higher frequency) structural alternative.



Verb Bias:

buy is biased towards DO

design is towards PD

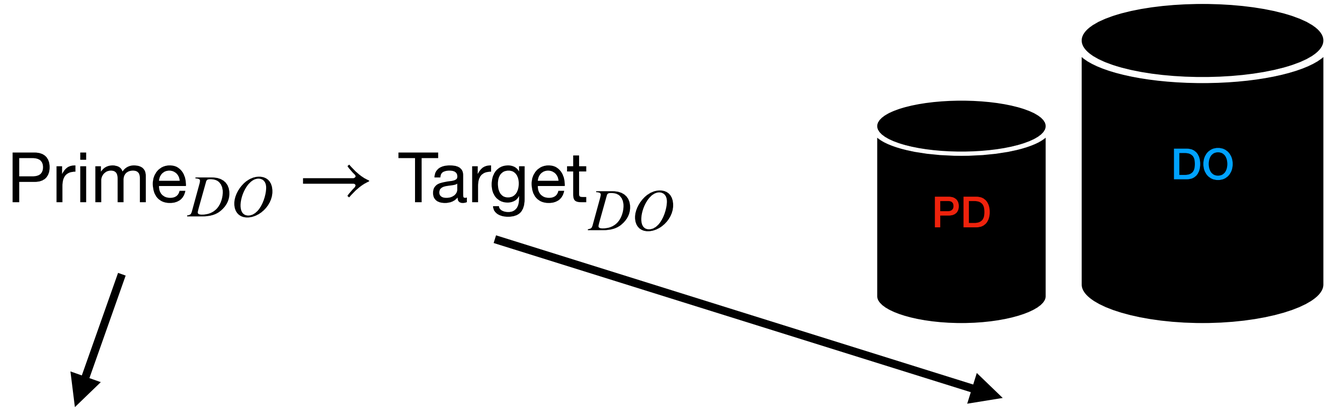
A doctor bought a chief a plate. … *The secretary drew the band the card.*

A doctor designed a chief a plate. … ***The secretary drew the band the card.***

[e.g., Ferreira 2003, Jaeger & Snider 2007, 2013]

The Inverse Frequency Effect (IFE)

- **Inverse Frequency Effect**: the *less preferred* (lower frequency) syntactic structure causes a *stronger* priming effect than the more preferred (higher frequency) structural alternative.



Verb Bias:

buy is biased towards DO

design is towards PD

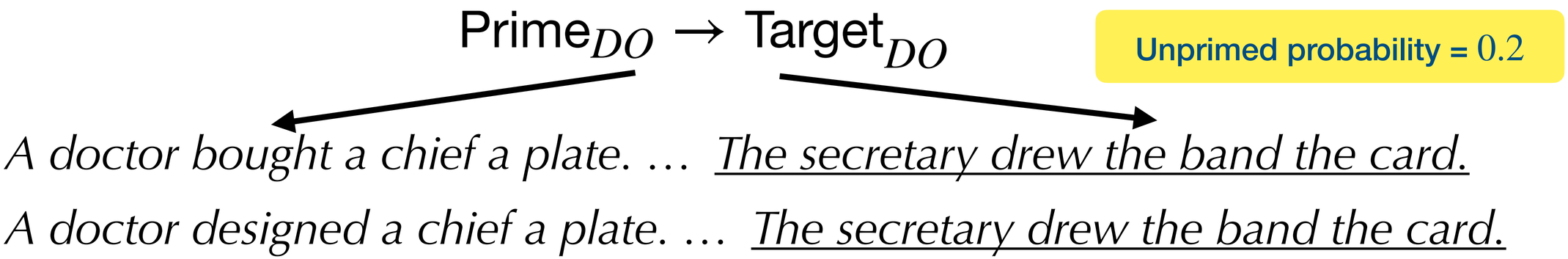
A doctor bought a chief a plate. … *The secretary drew the band the card.*

A doctor designed a chief a plate. … ***The secretary drew the band the card.***

Greater priming effect!

[e.g., Ferreira 2003, Jaeger & Snider 2007, 2013]

The Inverse Frequency Effect (IFE)



[e.g., Ferreira 2003, Jaeger & Snider 2007, 2013]

The Inverse Frequency Effect (IFE)

$\text{Prime}_{DO} \rightarrow \text{Target}_{DO}$

Unprimed probability = 0.2

A doctor bought a chief a plate. … *The secretary drew the band the card.*

A doctor designed a chief a plate. … *The secretary drew the band the card.*

Prime Verb

Bring
Buy
Find
Draw
Design

[e.g., Ferreira 2003, Jaeger & Snider 2007, 2013]

The Inverse Frequency Effect (IFE)

$\text{Prime}_{DO} \rightarrow \text{Target}_{DO}$

Unprimed probability = 0.2

A doctor bought a chief a plate. … *The secretary drew the band the card.*

A doctor designed a chief a plate. … *The secretary drew the band the card.*

Prime Verb	Verb PD Bias
Bring	0.23
Buy	0.27
Find	0.41
Draw	0.52
Design	0.77

[e.g., Ferreira 2003, Jaeger & Snider 2007, 2013]

The Inverse Frequency Effect (IFE)

$\text{Prime}_{DO} \rightarrow \text{Target}_{DO}$

Unprimed probability = 0.2

A doctor bought a chief a plate. … *The secretary drew the band the card.*

A doctor designed a chief a plate. … *The secretary drew the band the card.*

Prime Verb	Verb PD Bias	Primed probability (priming magnitude)
Bring	0.23	0.23 (0.03)
Buy	0.27	0.25 (0.05)
Find	0.41	0.28 (0.08)
Draw	0.52	0.30 (0.10)
Design	0.77	0.33 (0.13)

[e.g., Ferreira 2003, Jaeger & Snider 2007, 2013]

The Inverse Frequency Effect (IFE)

$\text{Prime}_{DO} \rightarrow \text{Target}_{DO}$

Unprimed probability = 0.2

A doctor bought a chief a plate. … *The secretary drew the band the card.*

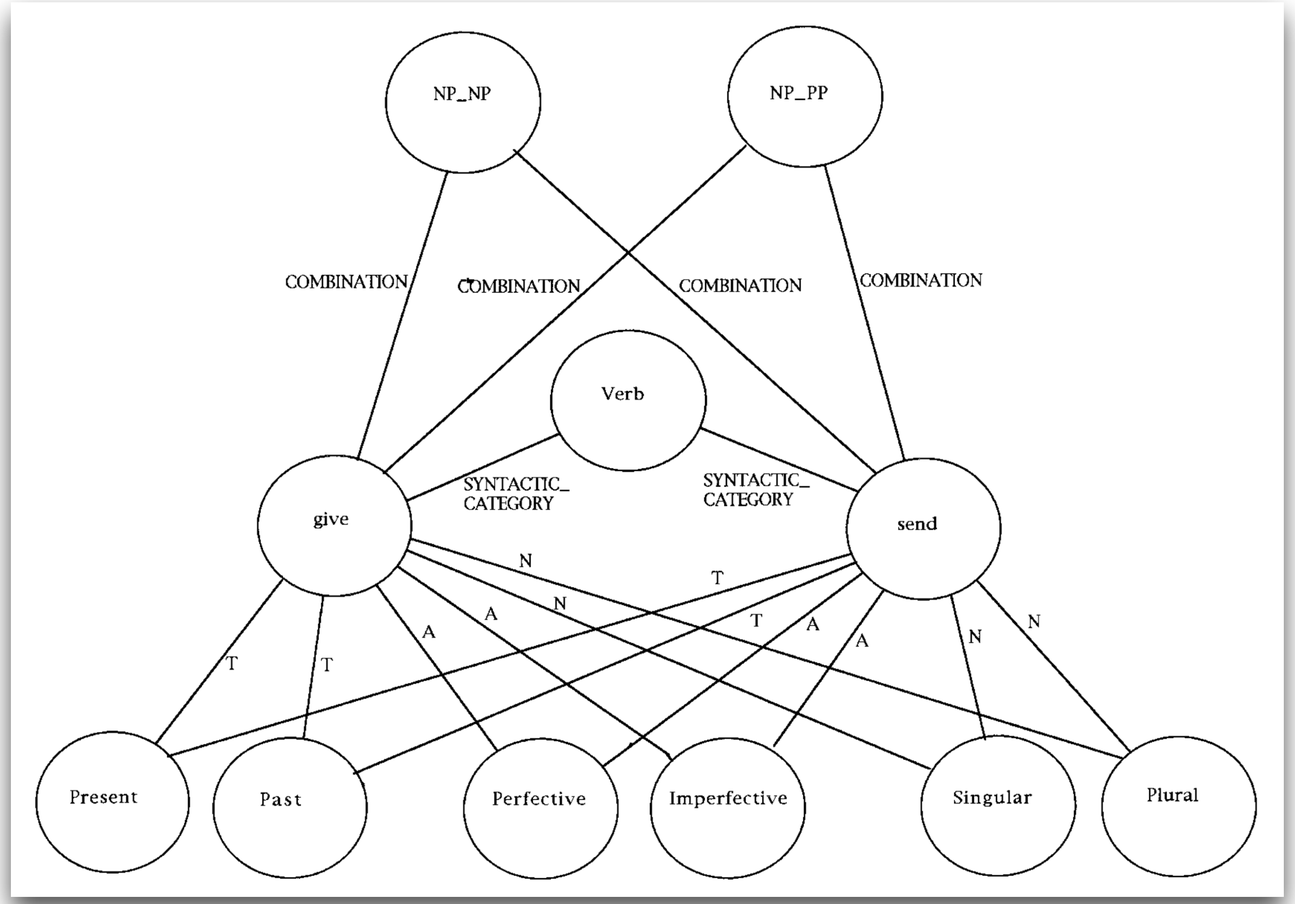
A doctor designed a chief a plate. … *The secretary drew the band the card.*

Prime Verb	Verb PD Bias	Primed probability (priming magnitude)
Bring	0.23	0.23 (0.03)
Buy	0.27	0.25 (0.05)
Find	0.41	0.28 (0.08)
Draw	0.52	0.30 (0.10)
Design	0.77	0.33 (0.13)

priming in DO *larger PD biases* *greater priming effects*

[e.g., Ferreira 2003, Jaeger & Snider 2007, 2013]

Theory 1: Transient Activation

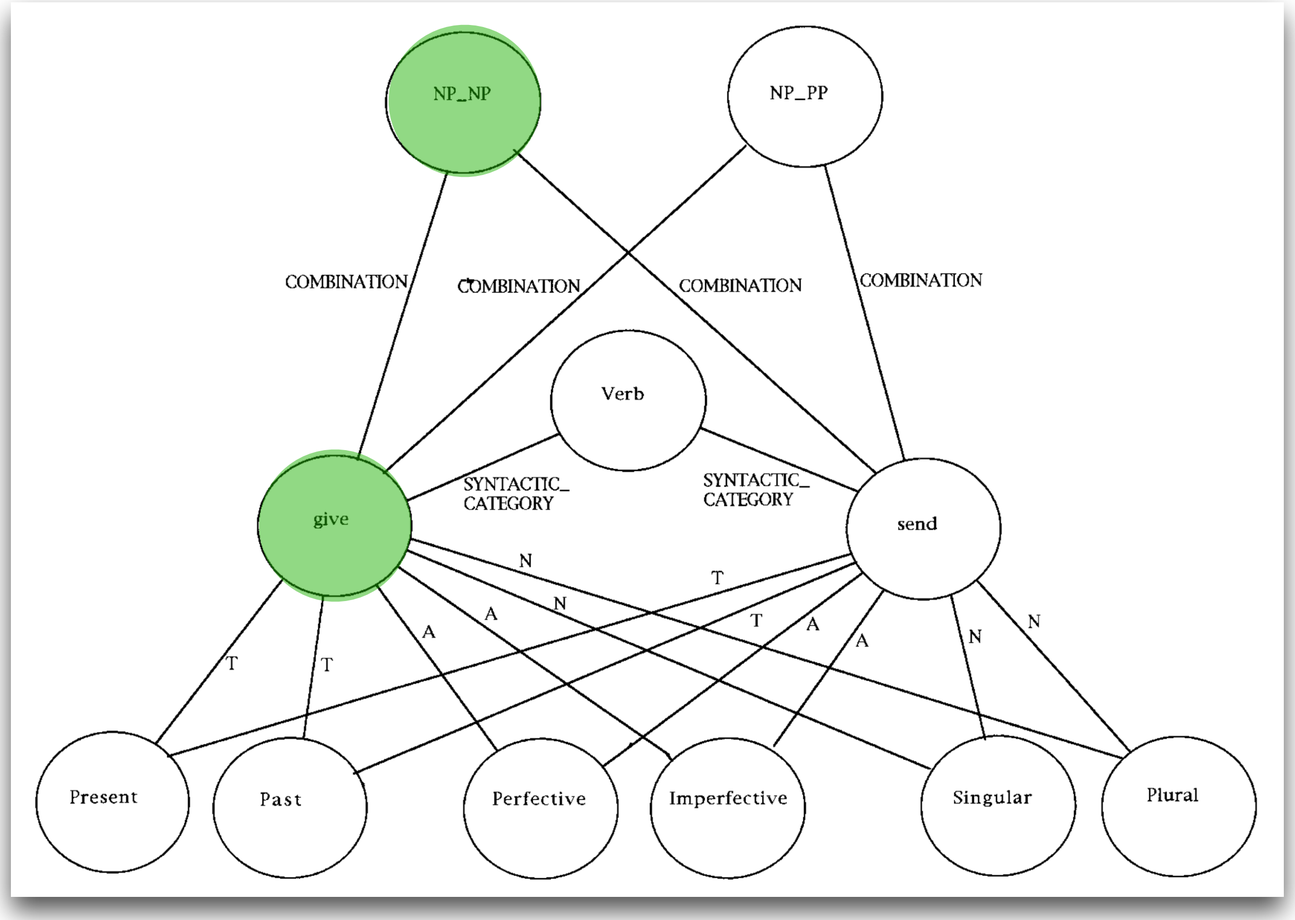


← Structural Nodes

← Lexical Nodes

[e.g., Pickering & Branigan 1998]

Theory 1: Transient Activation



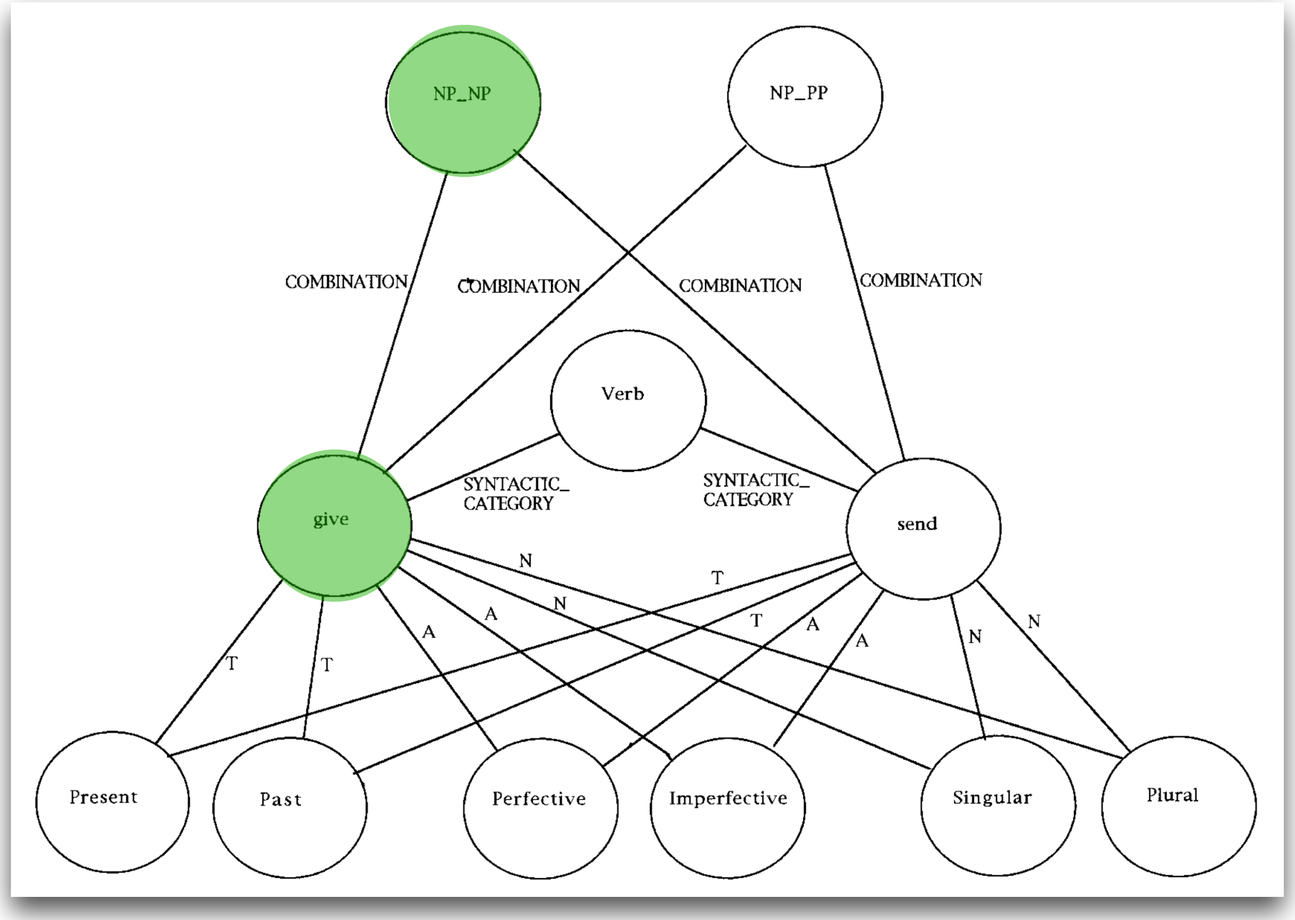
Structural Nodes

Lexical Nodes

[e.g., Pickering & Branigan 1998]

Theory 1: Transient Activation

Quickly decaying activations



← Structural Nodes

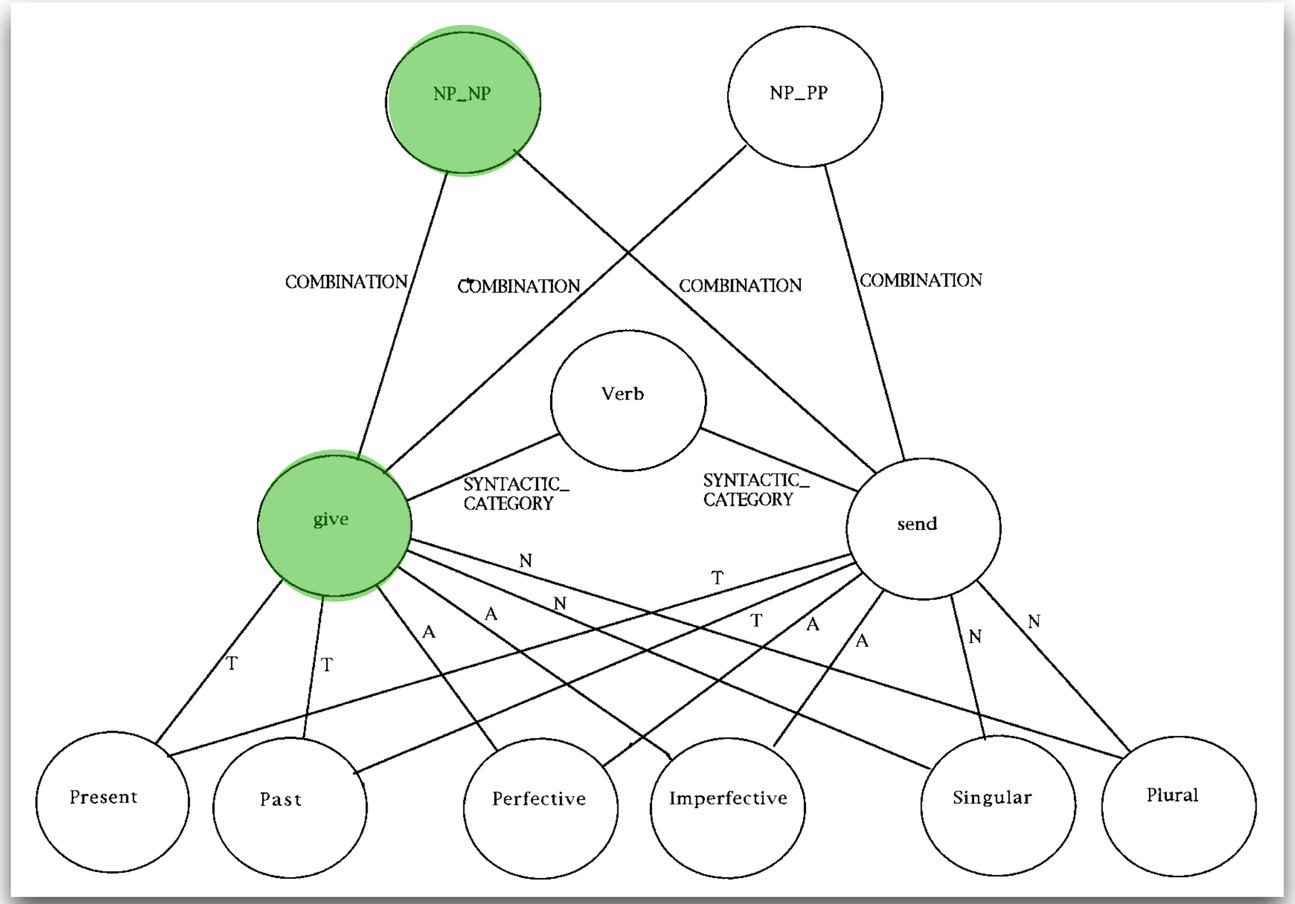
Increasing the activation values of the nodes, which persists through next productions.

← Lexical Nodes

[e.g., Pickering & Branigan 1998]

Theory 1: Transient Activation

Quickly decaying activations



← Structural Nodes

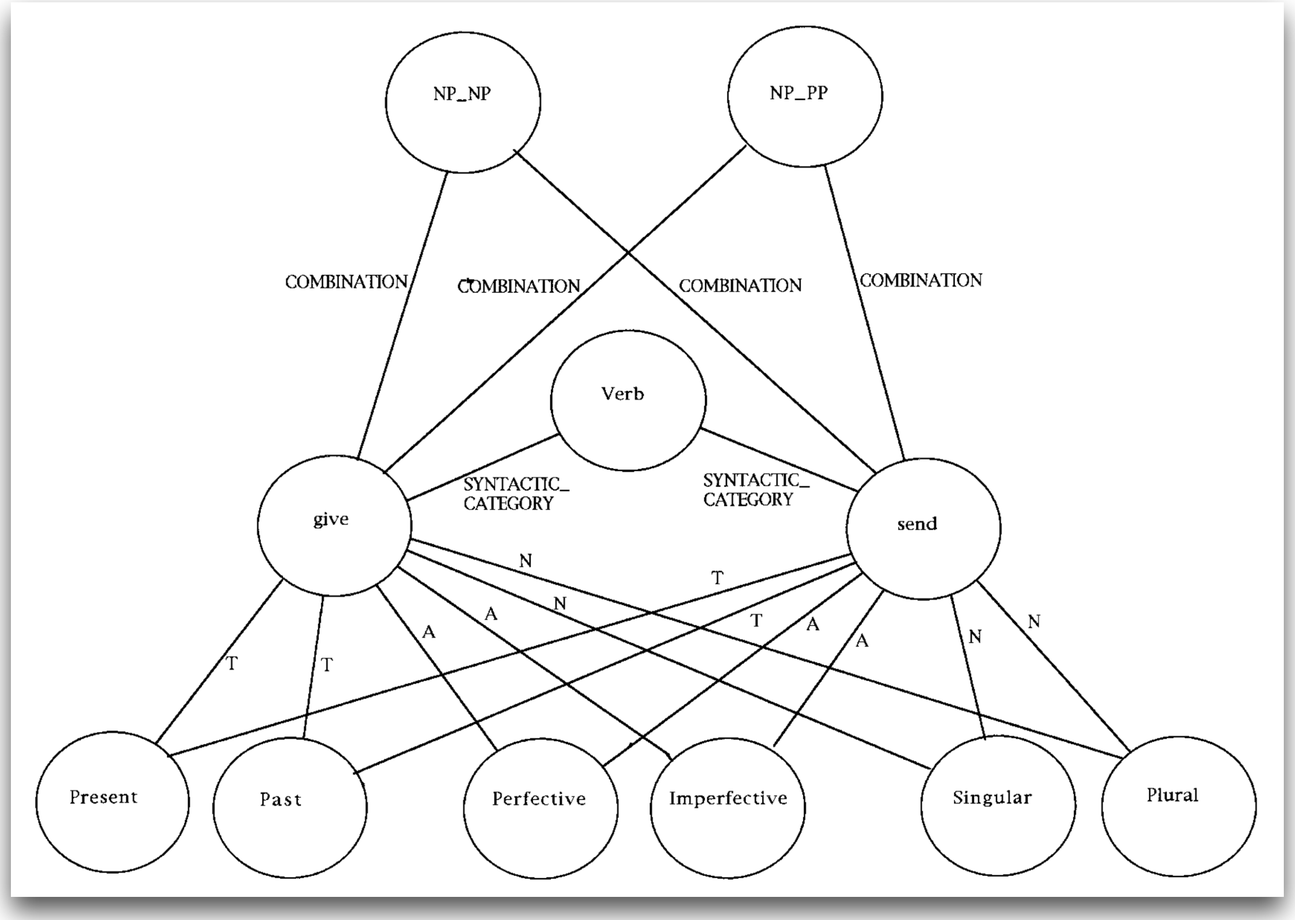
Increasing the activation values of the nodes, which persists through next productions.

← Lexical Nodes

- An activation-based mechanism
- Explaining short-term effect
- Inverse frequency effect? ❌

[e.g., Pickering & Branigan 1998]

Theory 2: Implicit Learning

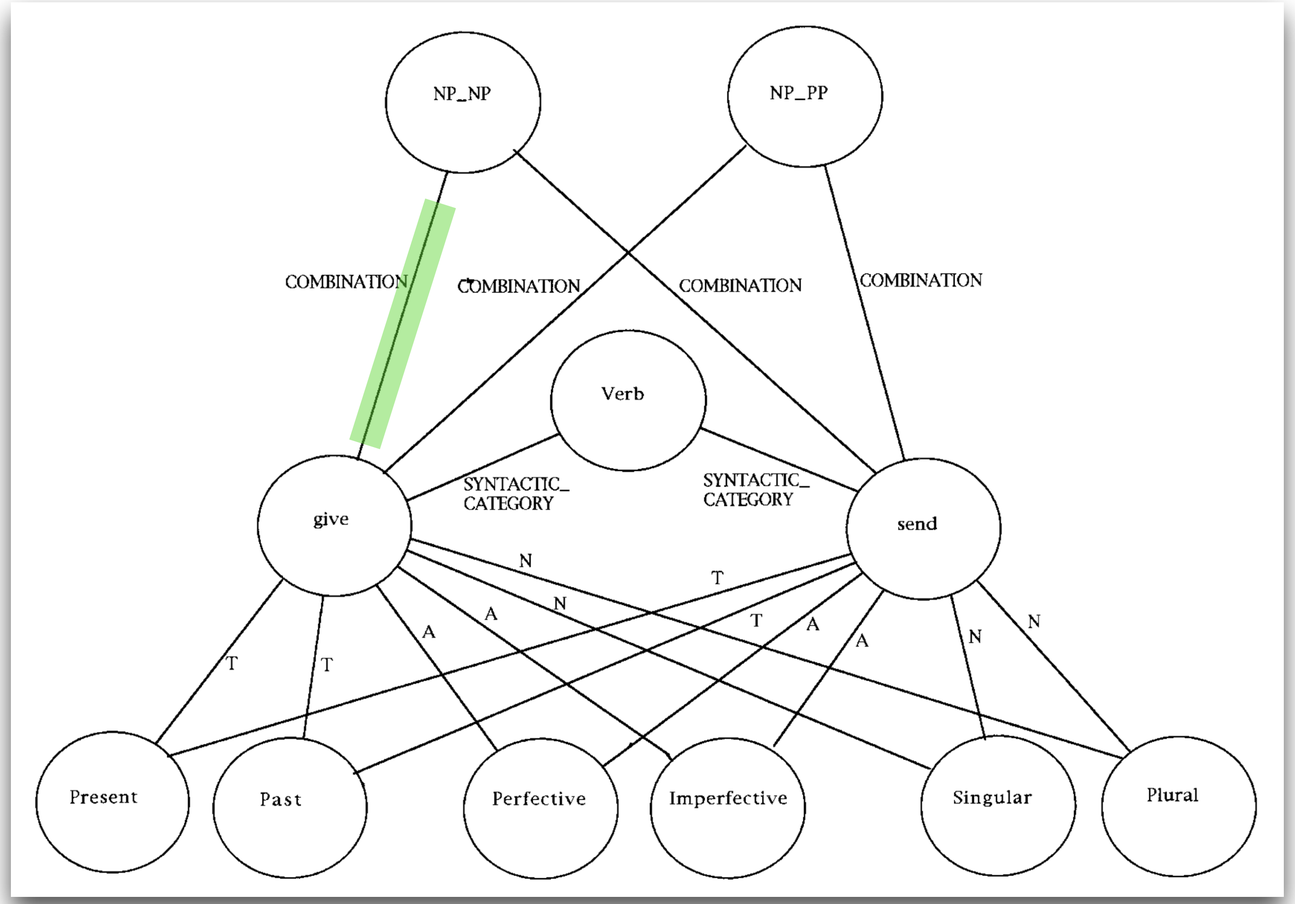


← Structural Nodes

← Lexical Nodes

[e.g., Chang et al. 2002, 2006; Jaeger & Snider 2013]

Theory 2: Implicit Learning



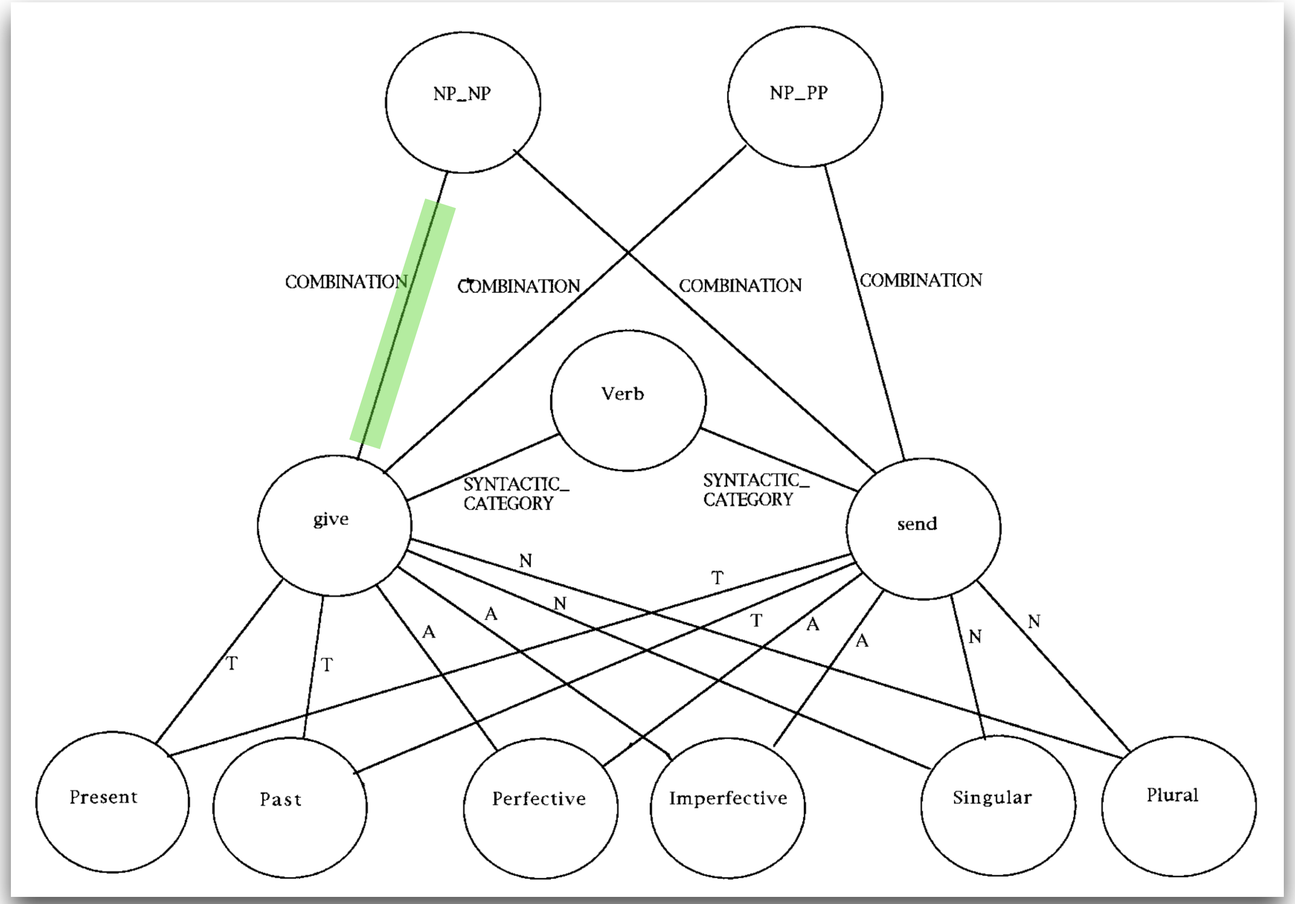
← Structural Nodes

← Lexical Nodes

[e.g., Chang et al. 2002, 2006; Jaeger & Snider 2013]

Theory 2: Implicit Learning

Slowly decaying weight changes



← Structural Nodes

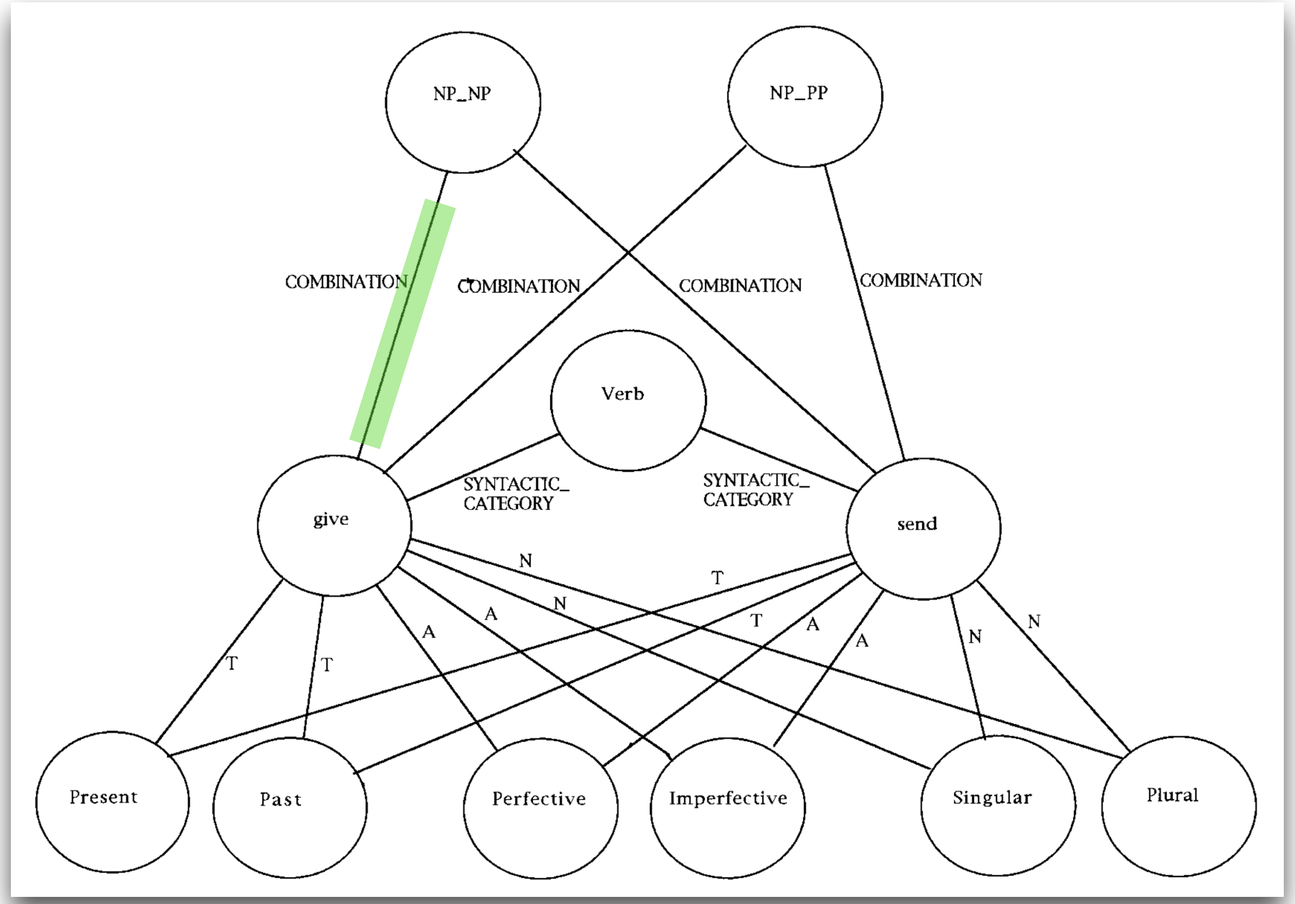
Increasing the link weights proportional to the error signal between verb-bias prediction and the actual input prime.

← Lexical Nodes

[e.g., Chang et al. 2002, 2006; Jaeger & Snider 2013]

Theory 2: Implicit Learning

Slowly decaying weight changes



← Structural Nodes

Increasing the link weights <u>proportional to</u> the error signal between verb-bias prediction and the actual input prime.

← Lexical Nodes

- An error-driven learning / adaptation mechanism
- Explaining long-term effect
- Inverse Frequency effect? ✅

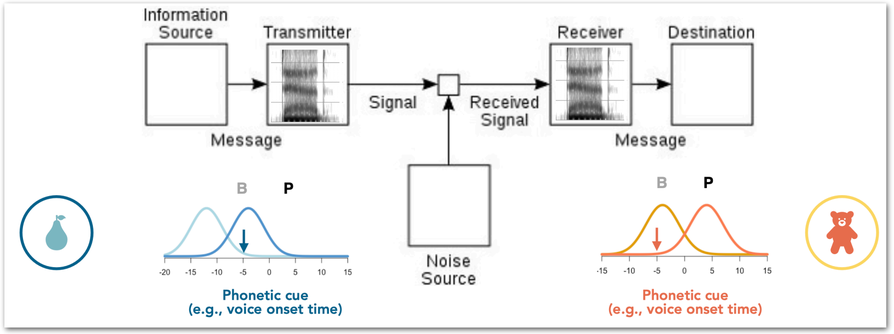
[e.g., Chang et al. 2002, 2006; Jaeger & Snider 2013]

Side Notes: Priming as Linguistic Adaptation

[e.g., Xie & Kurumada 2023; Lu et al. 2025; Feng 2022; etc.]

Side Notes: Priming as Linguistic Adaptation

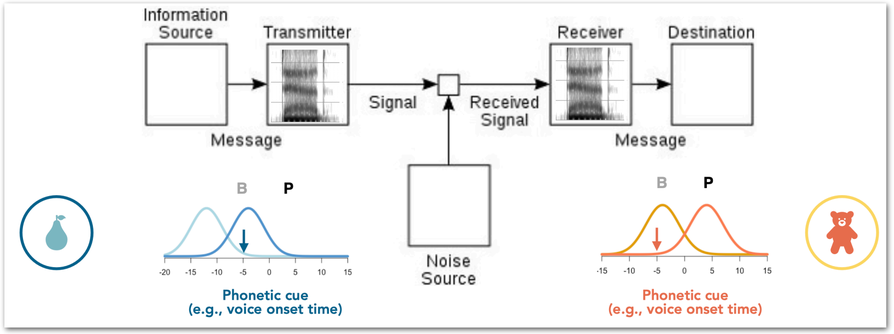
- **Linguistic Adaptation:** adjustment of distributional knowledge / expectations to accommodate local input;
  - Speech adaptation & perception;
  - Syntactic satiation;
  - Informativeness calibration & presupposition/pragmatic routine calibration;



[e.g., Xie & Kurumada 2023; Lu et al. 2025; Feng 2022; etc.]

Side Notes: Priming as Linguistic Adaptation

- **Linguistic Adaptation:** adjustment of distributional knowledge / expectations to accommodate local input;
  - Speech adaptation & perception;
  - Syntactic satiation;
  - Informativeness calibration & presupposition/pragmatic routine calibration;

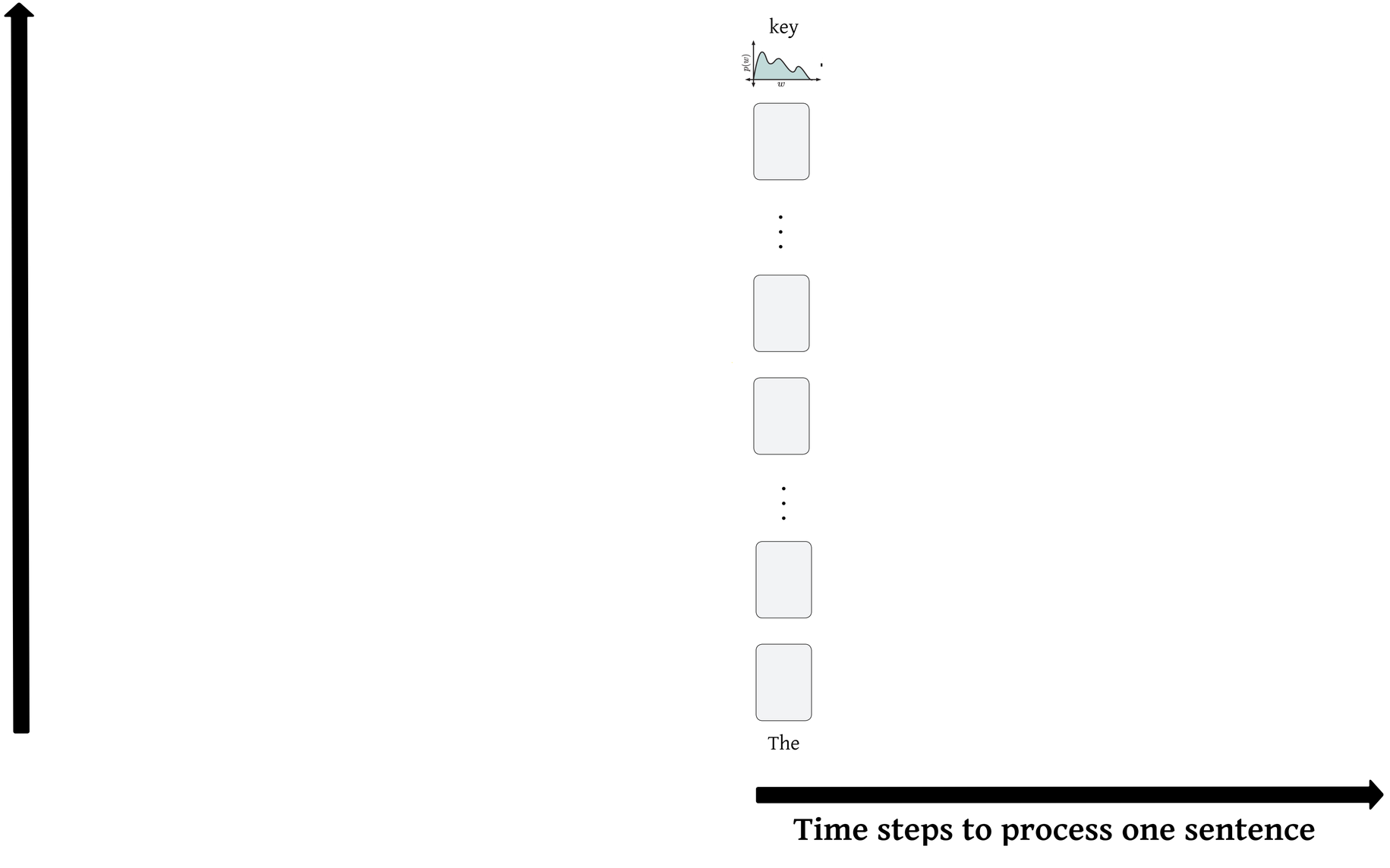


- **Predictive coding gives rise to linguistic adaptation**
  - Predictive coding: the system constantly predicts (linguistic) input and update the mental model based on the mismatch between predicted vs. actual input.
  - **The IFE is the *signature* that the update is error-driven**, not just residual activation: rarer structures create bigger prediction error → bigger update.

[e.g., Xie & Kurumada 2023; Lu et al. 2025; Feng 2022; etc.]

In-Context Learning in LLMs

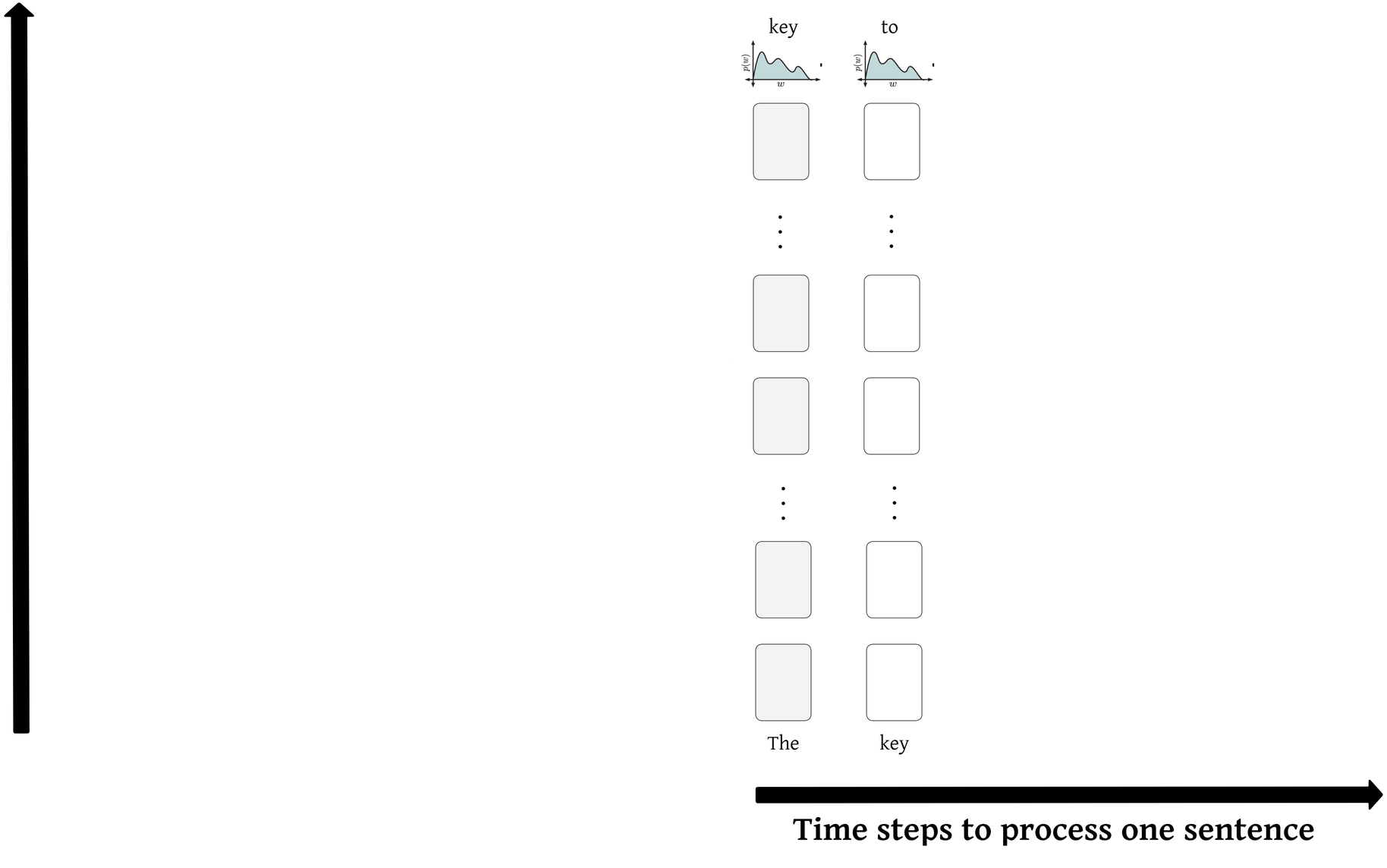
A Primer on Intuitions behind LLMs



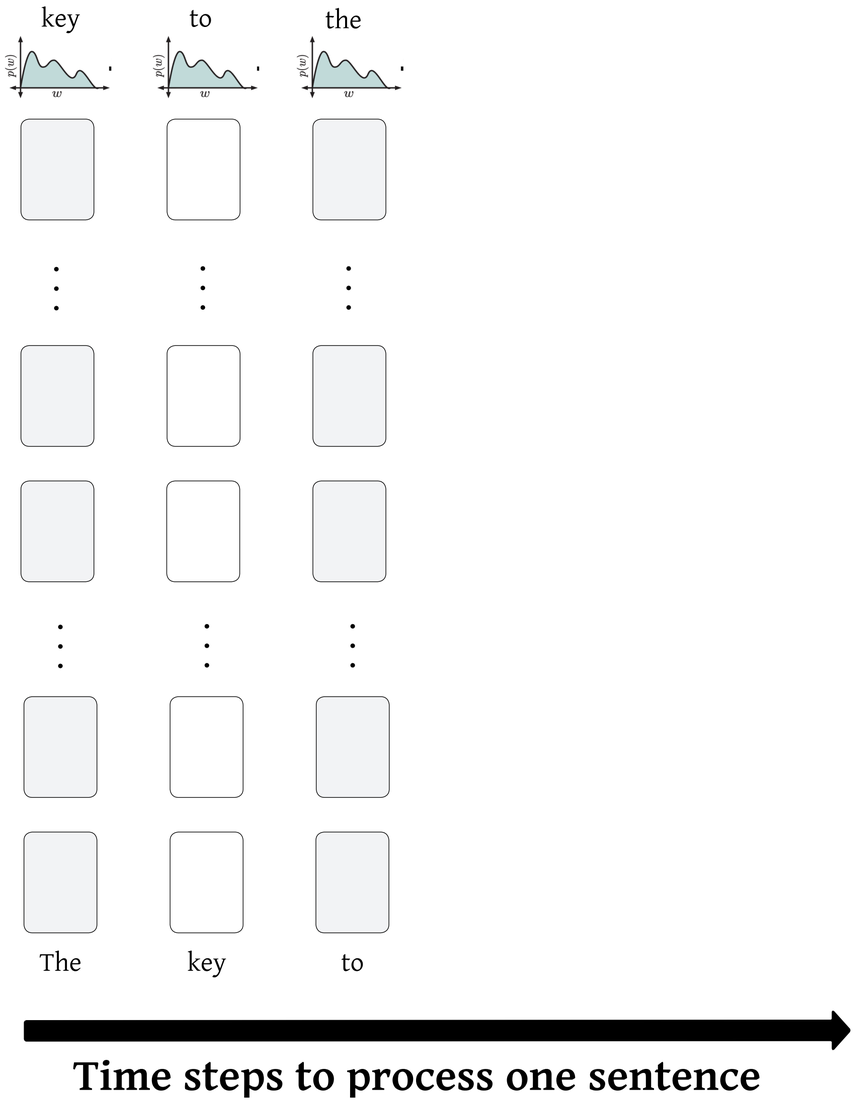
A Primer on Intuitions behind LLMs

Time steps to process one sentence

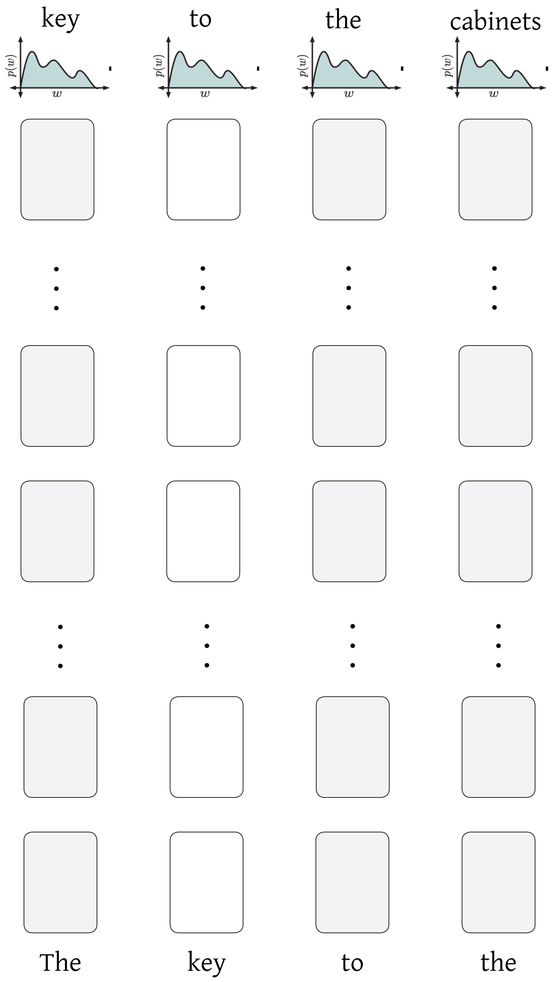
A Primer on Intuitions behind LLMs



A Primer on Intuitions behind LLMs

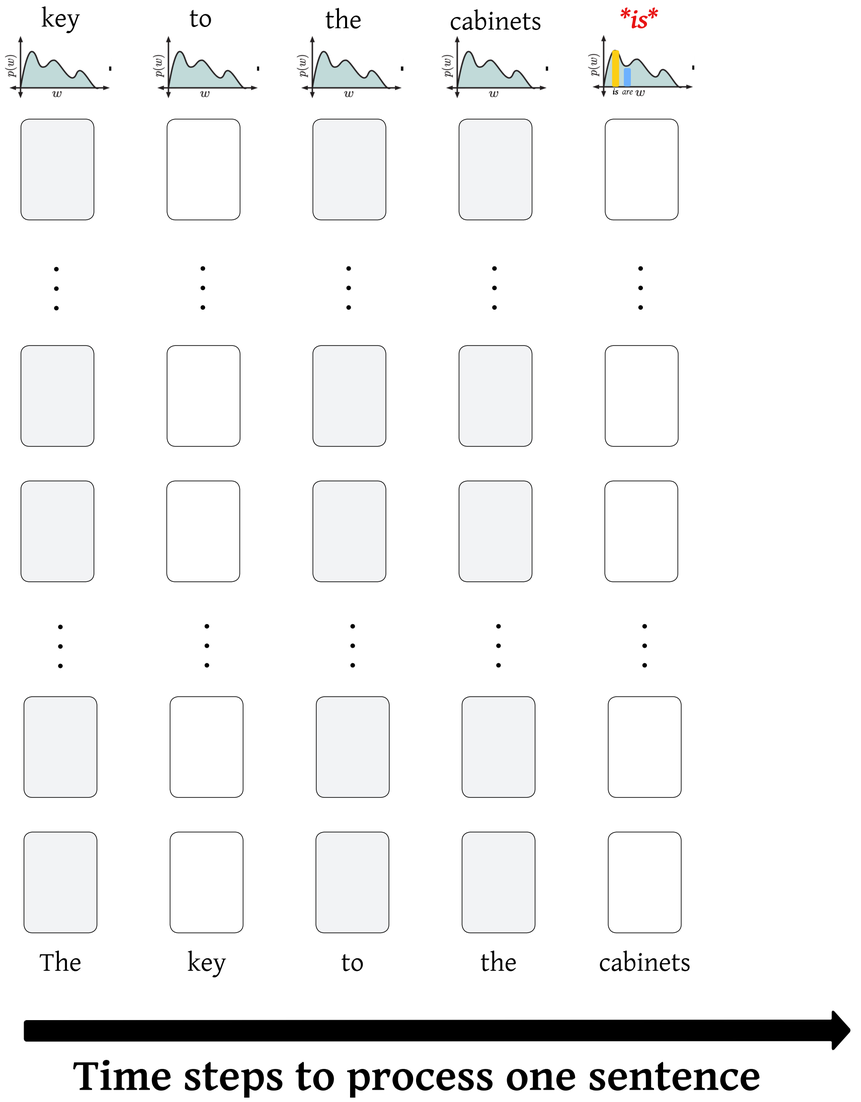


A Primer on Intuitions behind LLMs

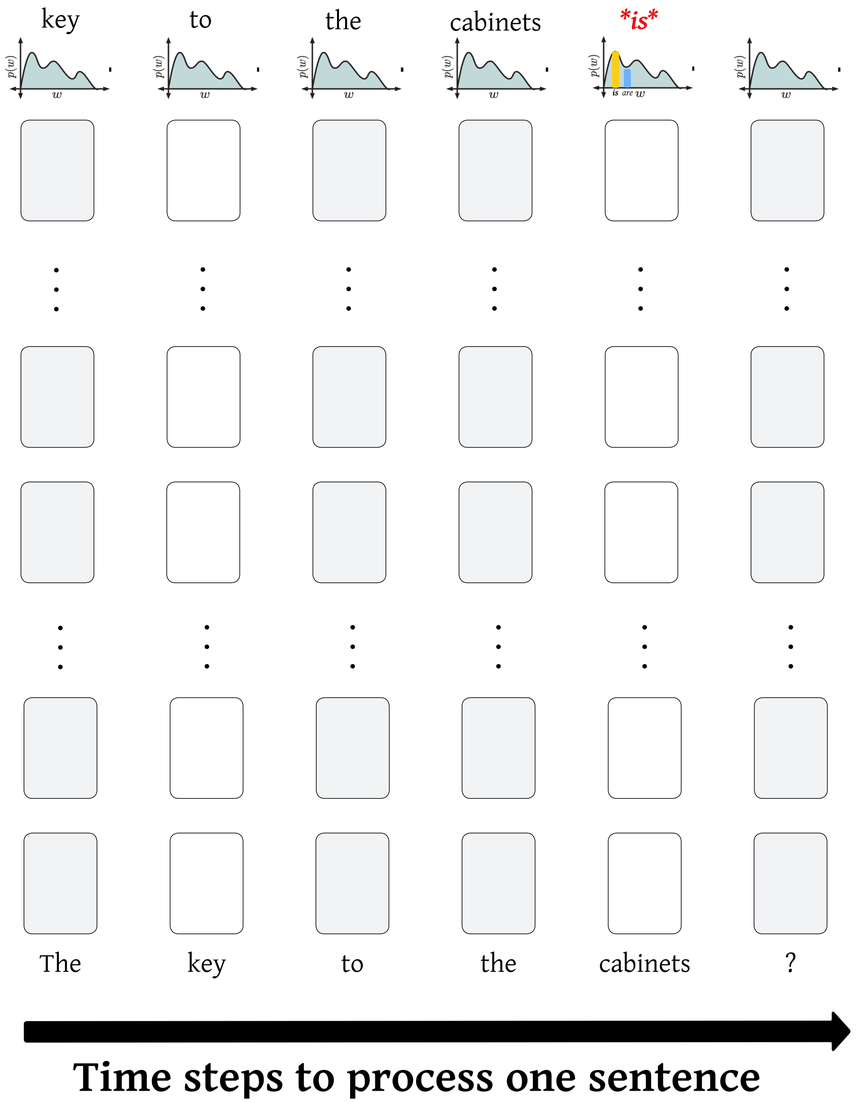


Time steps to process one sentence

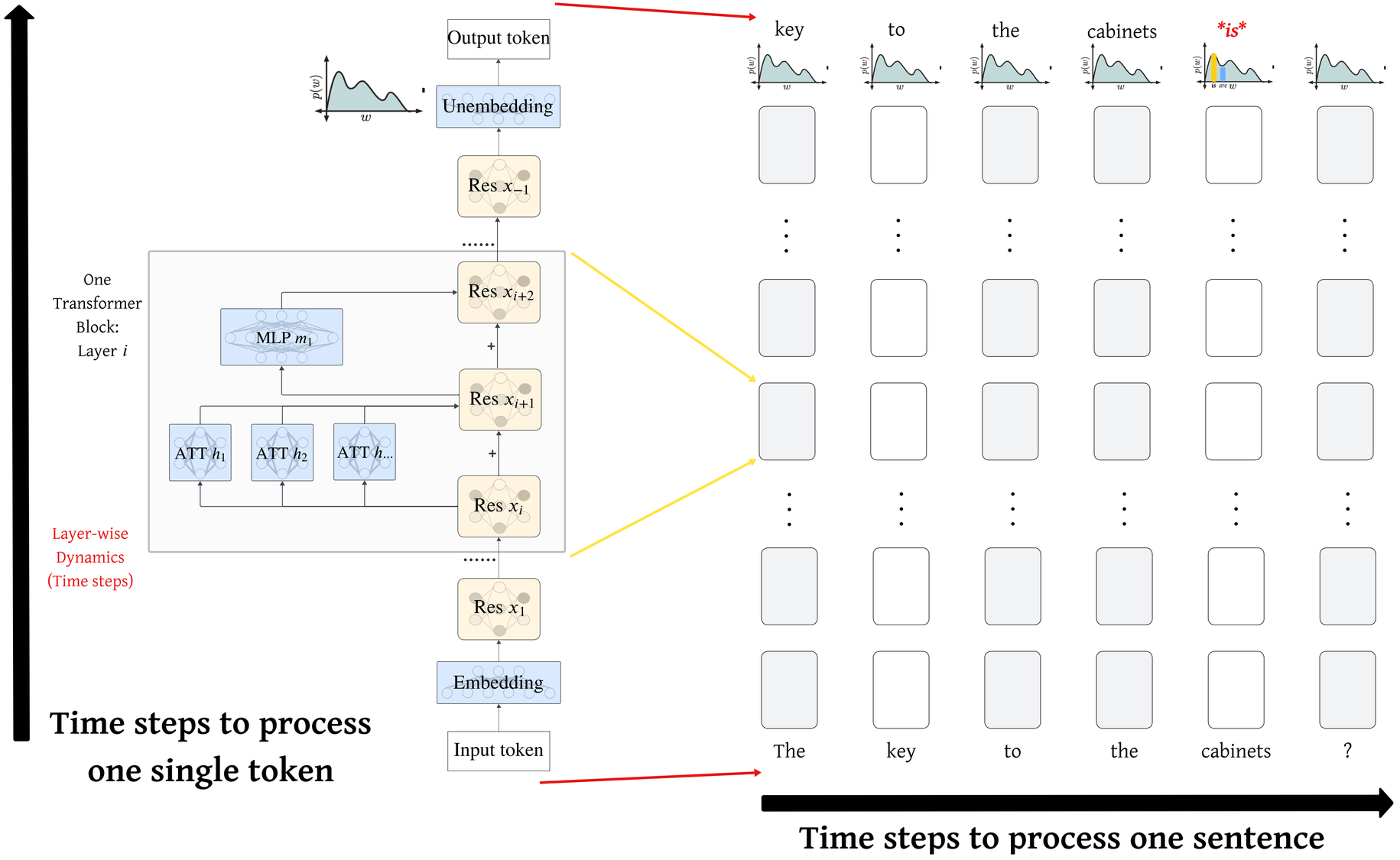
A Primer on Intuitions behind LLMs



A Primer on Intuitions behind LLMs



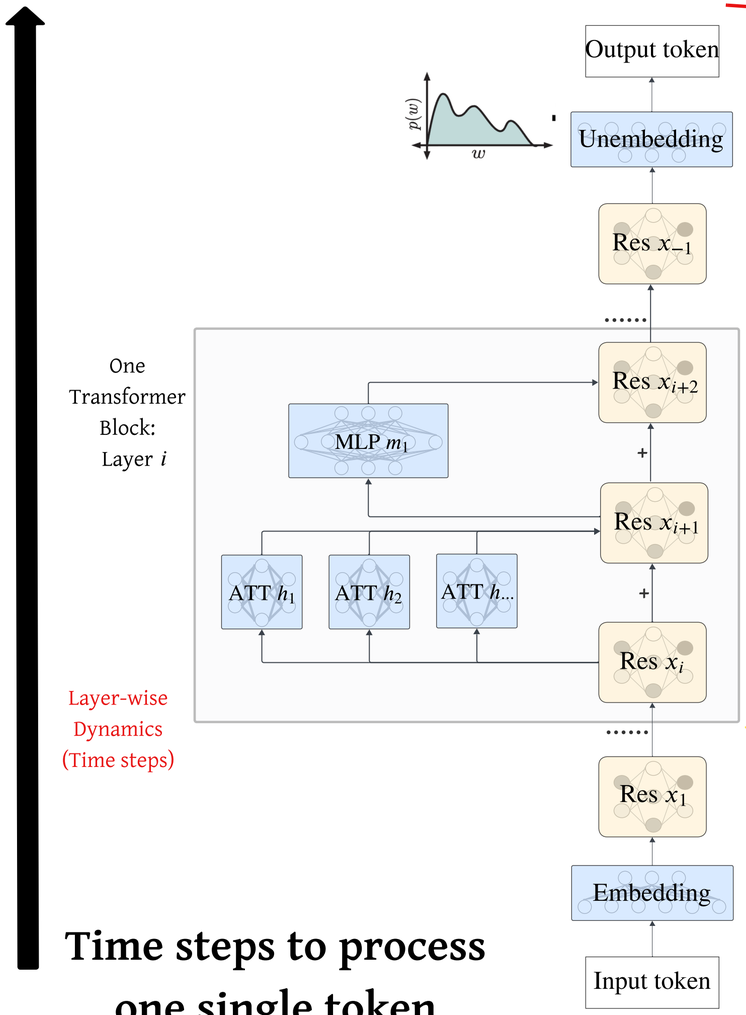
A Primer on Intuitions behind LLMs

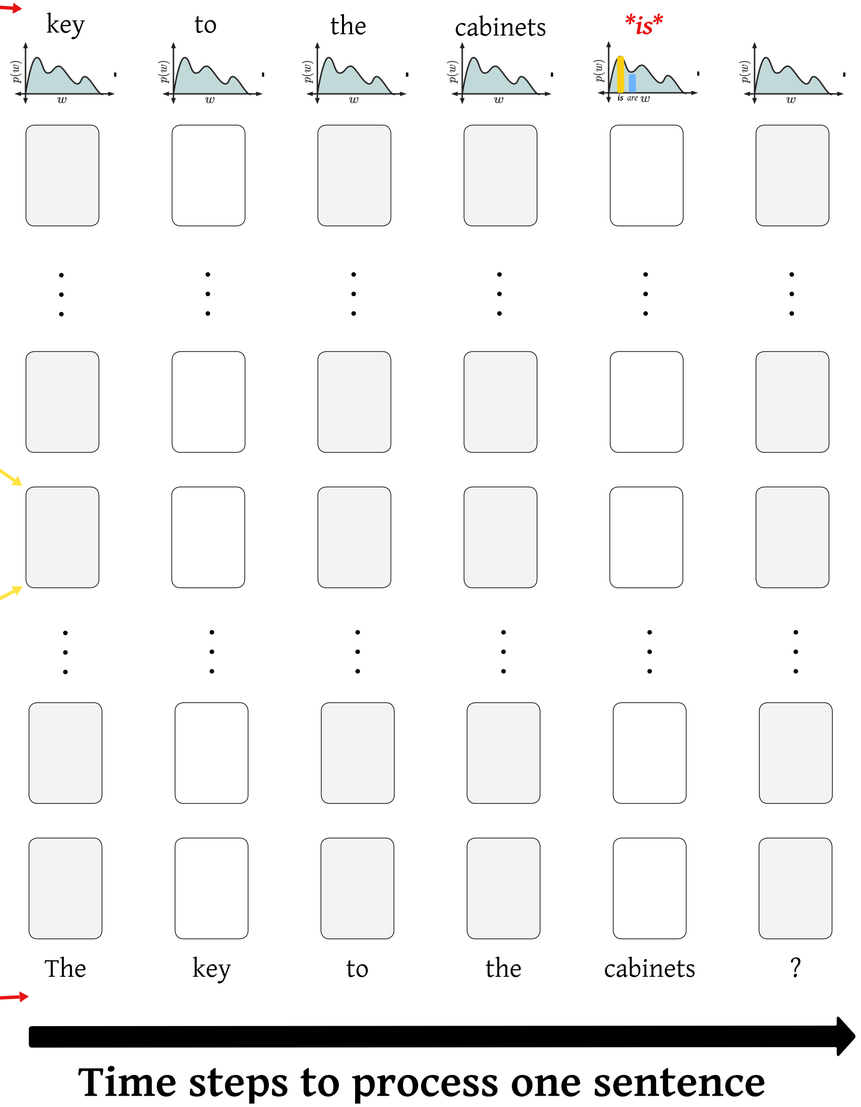


A Primer on Intuitions behind LLMs

Modern LLMs: machines of *next-token prediction*

- **Parameters:** trainable weights updated during training & fine-tuning → long-term memory-ish;
- **Activations:** temporary variables generarted during running → working memory;





In-Context Learning (ICL) in LLMs

In-Context Learning: several demonstration-answer pairs of a task in context improves models' performance on this task.

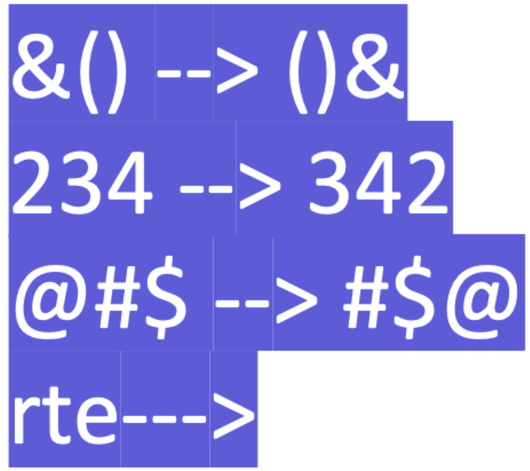
- No explicit description of the task;
- No weight (long-term memory) update!

[Brown et al. 2020]

In-Context Learning (ICL) in LLMs

In-Context Learning: several demonstration-answer pairs of a task in context improves models' performance on this task.

- No explicit description of the task;
- No weight (long-term memory) update!

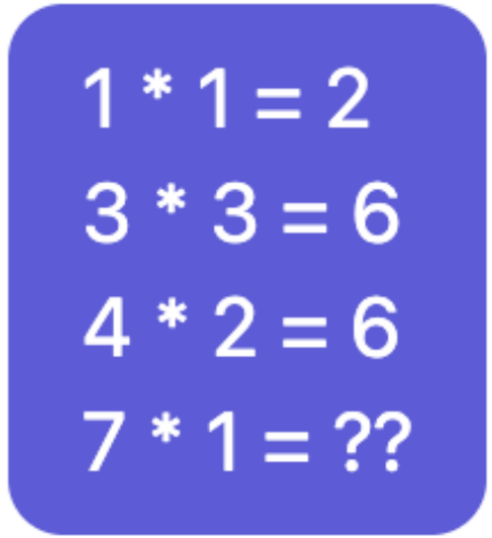


[Brown et al. 2020]

In-Context Learning (ICL) in LLMs

In-Context Learning: several demonstration-answer pairs of a task in context improves models' performance on this task.

- No explicit description of the task;
- No weight (long-term memory) update!

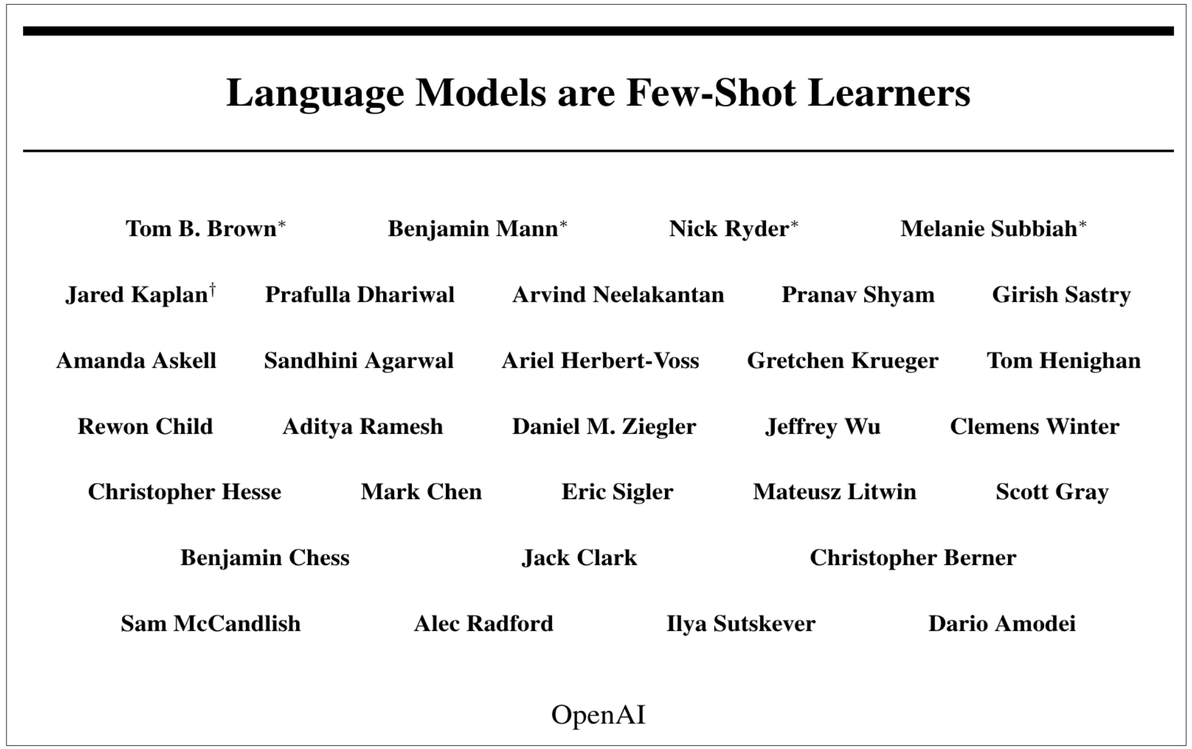


[Brown et al. 2020]

In-Context Learning (ICL) in LLMs

In-Context Learning: several demonstration-answer pairs of a task in context improves models' performance on this task.

- No explicit description of the task;
- No weight (long-term memory) update!



Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

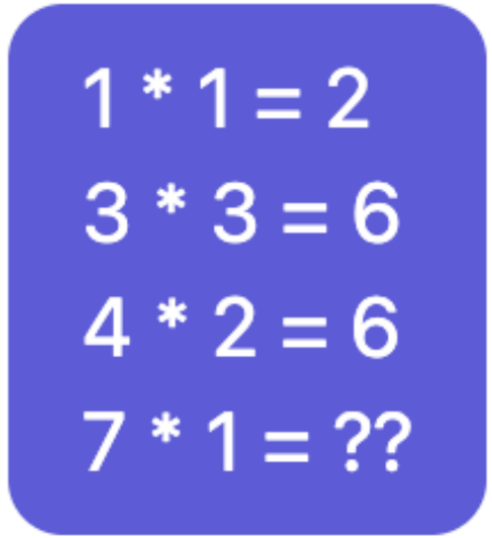
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI



[Brown et al. 2020]

ICL vs. In-*Weights* Learning

[Brown et al. 2020]

[Brown et al. 2020]

ICL vs. In-*Weights* Learning

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

[Brown et al. 2020]

In-context Learning:

- No gradient updates;
- Rapid: from a few examples;
- One-shot / Few-shot learning;

[Brown et al. 2020]

ICL vs. In-*Weights* Learning

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

[Brown et al. 2020]

In-context Learning:

- No gradient updates;
- Rapid: from a few examples;
- One-shot / Few-shot learning;

In-weights Learning (Fine-tuning):

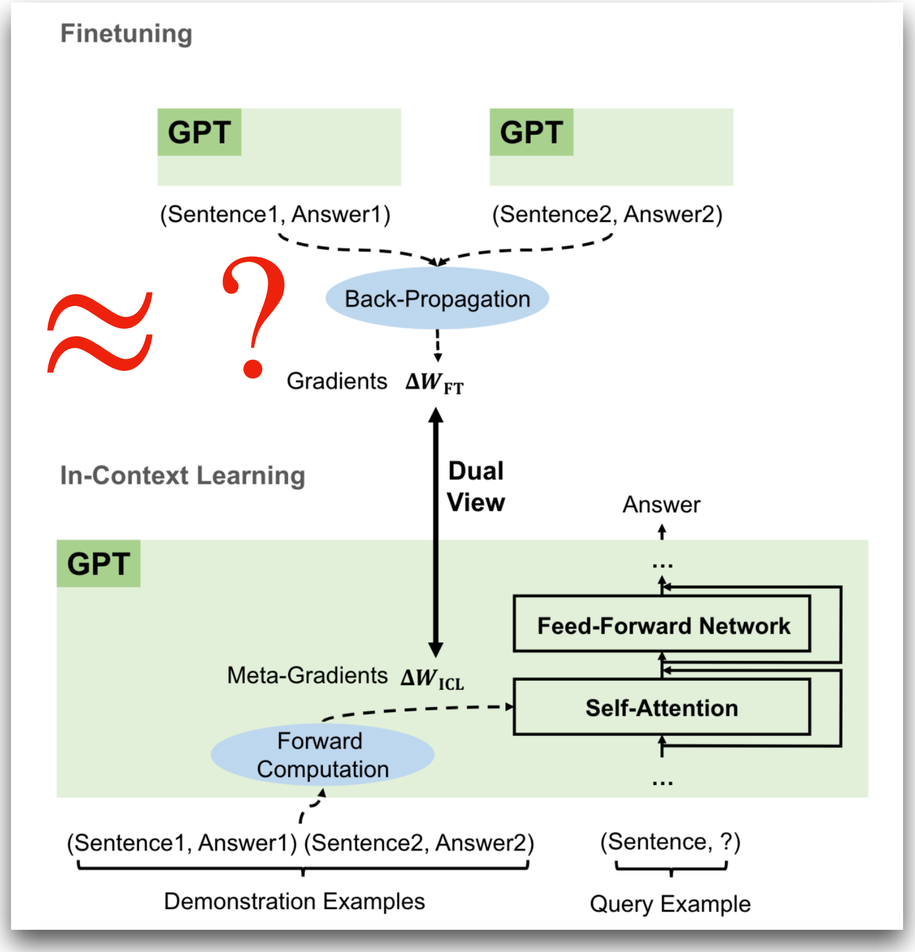
- Gradient-based;
- Slow: need many examples;
- Standard supervised learning;

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

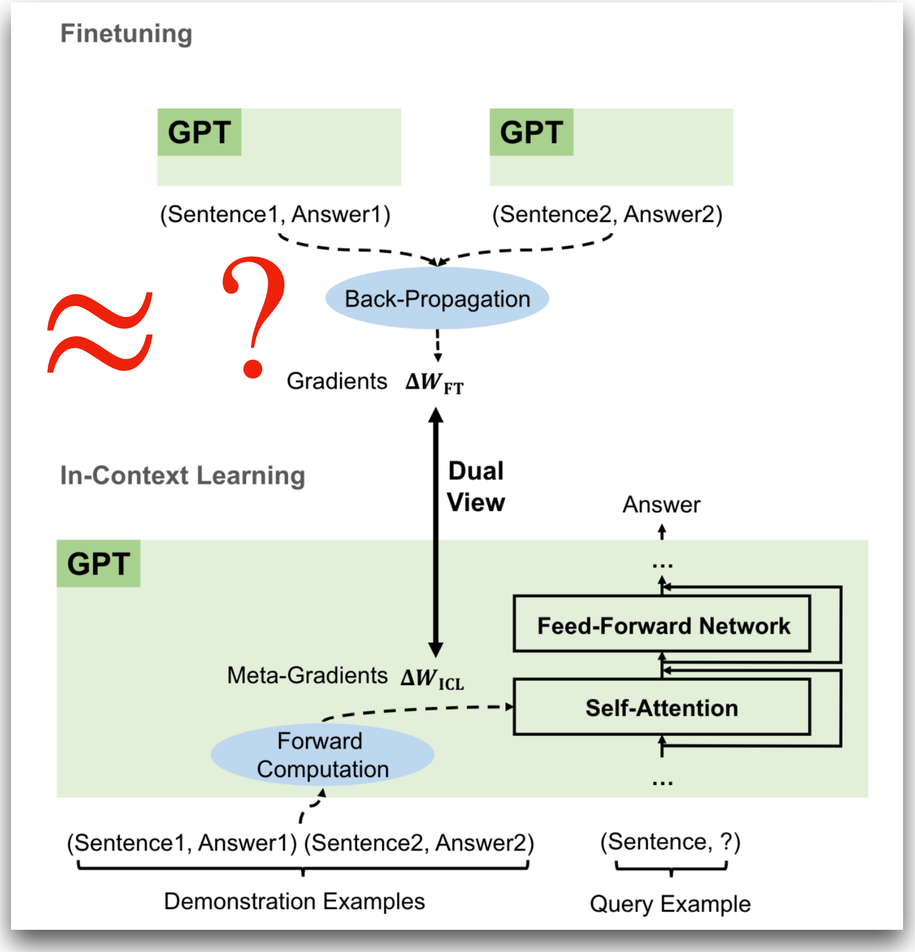
[Brown et al. 2020]

Research Question: is there error-driven learning mechanisms in ICL?



[e.g., Xie et al. 2022, von Oswald et al. 2022, Dai et al. 2023]

Research Question: is there error-driven learning mechanisms in ICL?

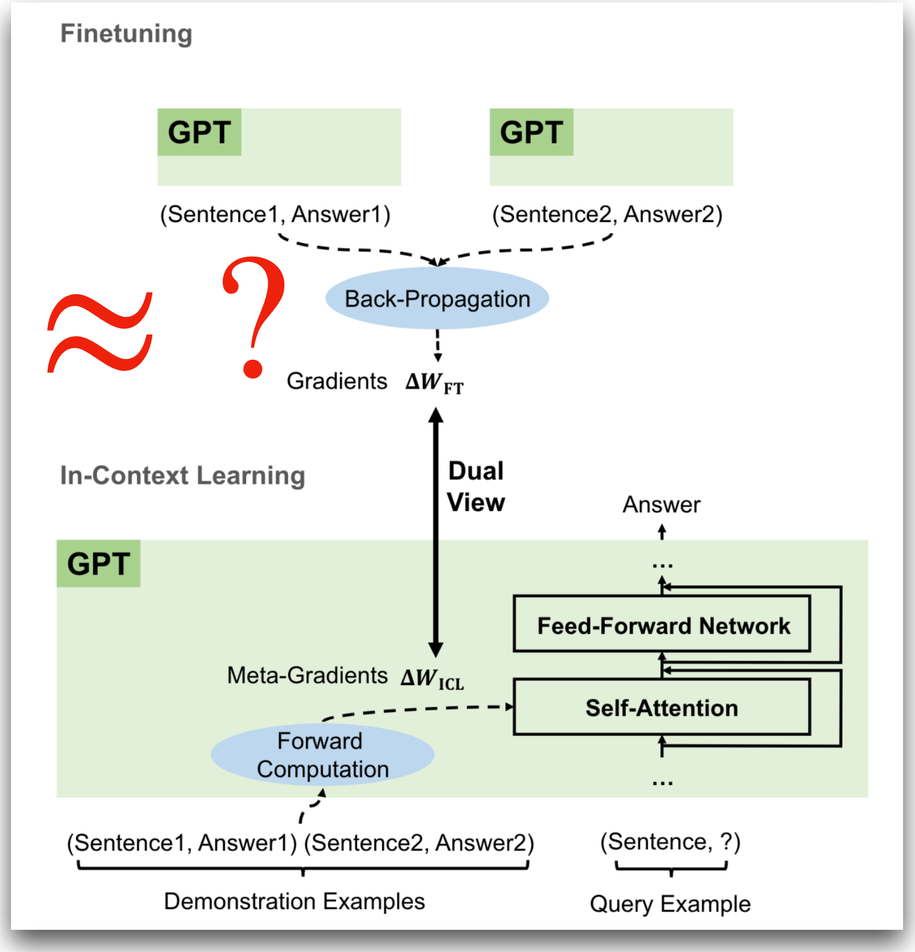


Previous claims: ICL may be implicitly doing learning

- Implicit Bayesian update; gradient descent; meta-optimization / implicit fine-tuning….

[e.g., Xie et al. 2022, von Oswald et al. 2022, Dai et al. 2023]

Research Question: is there error-driven learning mechanisms in ICL?



Previous claims: ICL may be implicitly doing learning

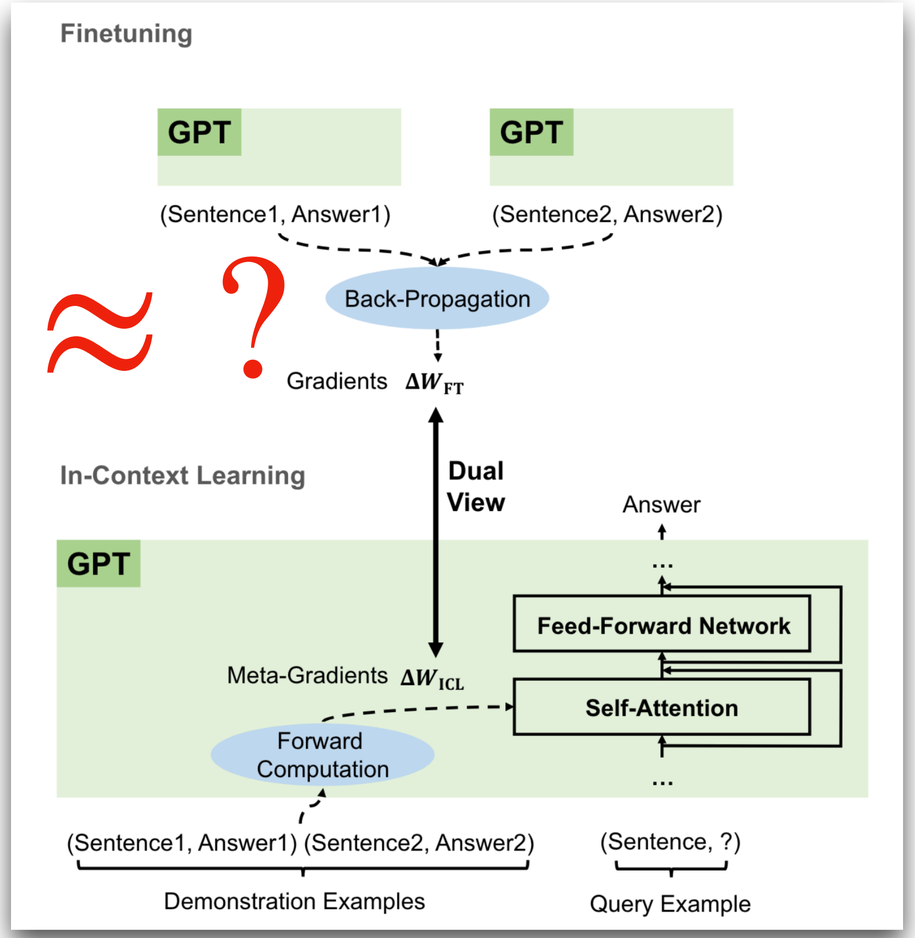
- Implicit Bayesian update; gradient descent; meta-optimization / implicit fine-tuning….

Previous studies in AI/CS:

- On toy models and toy tasks;
- In-principle claims: math derivations with assumptions on architecture;

[e.g., Xie et al. 2022, von Oswald et al. 2022, Dai et al. 2023]

Research Question: is there error-driven learning mechanisms in ICL?



Previous claims: ICL may be implicitly doing learning

- Implicit Bayesian update; gradient descent; meta-optimization / implicit fine-tuning….

Previous studies in AI/CS:

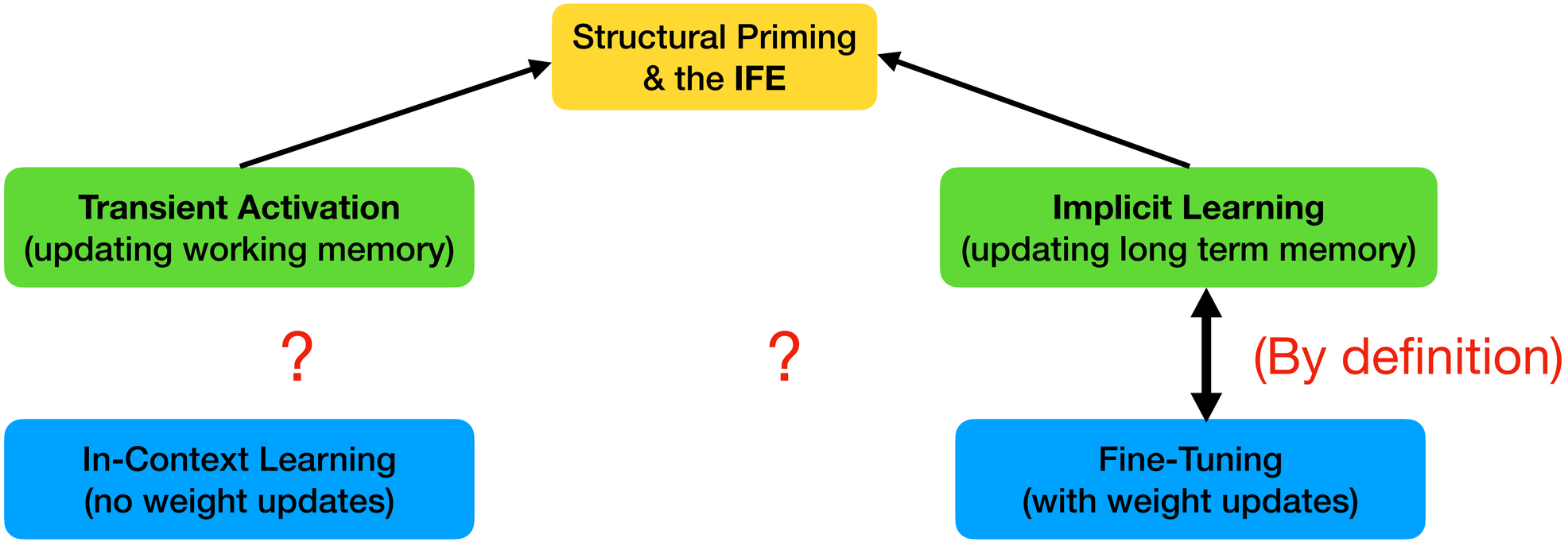
- On toy models and toy tasks;
- In-principle claims: math derivations with assumptions on architecture;

Current work:

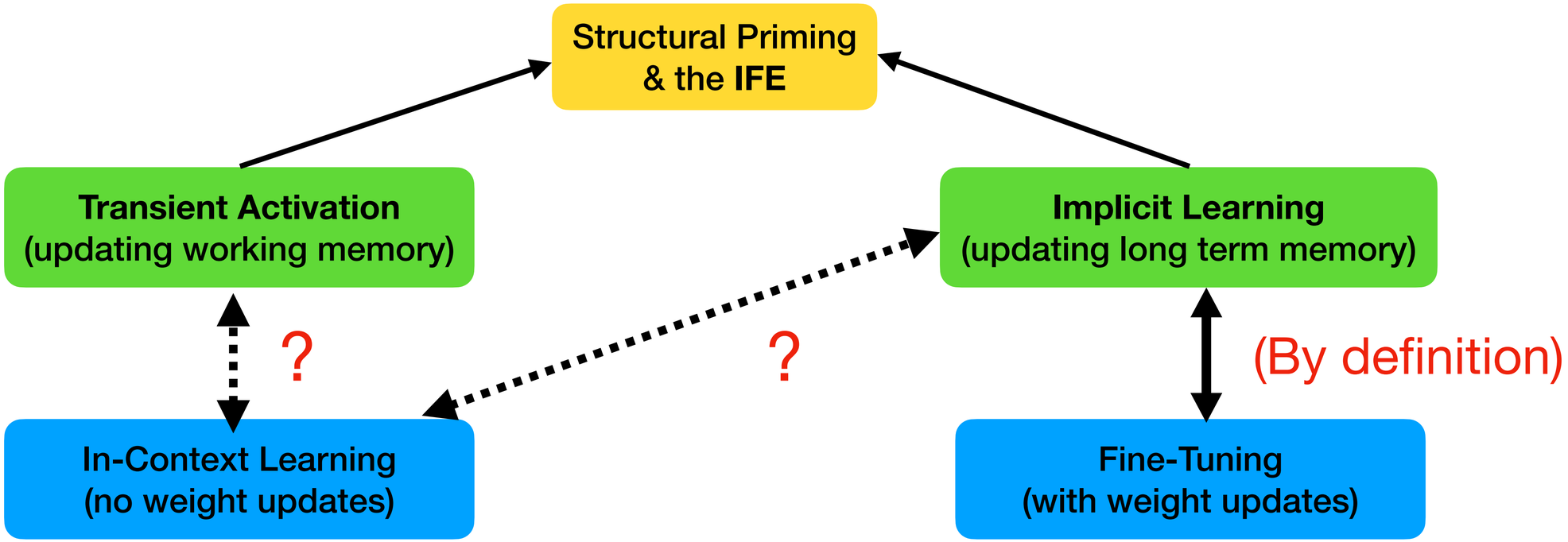
- On pretrained, actual LLMs;
- Psycholinguistically motivated, more testing a more fine-grained effect.

[e.g., Xie et al. 2022, von Oswald et al. 2022, Dai et al. 2023]

Current Picture

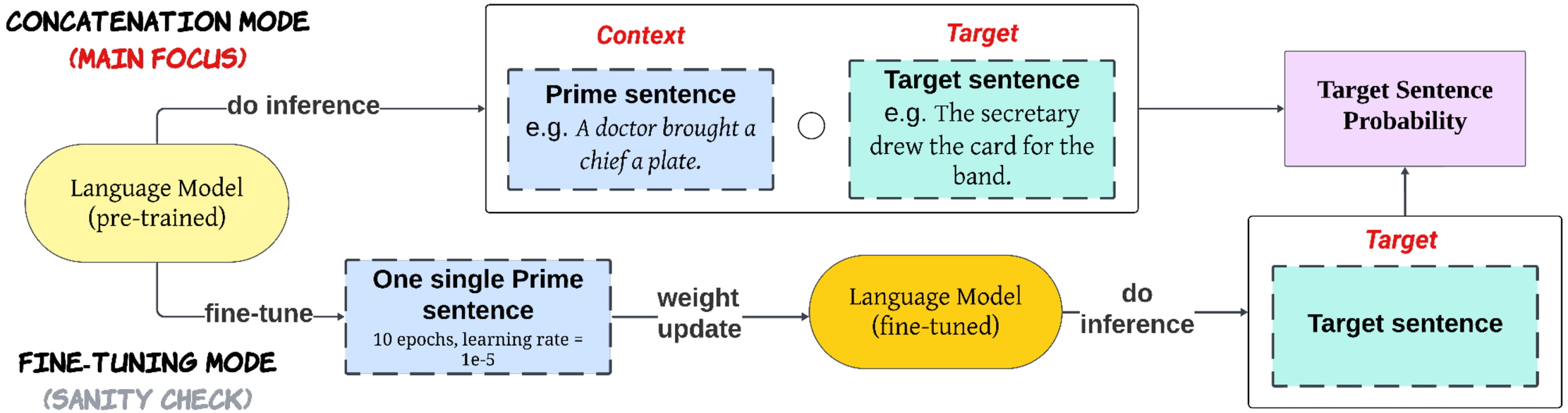


Current Picture



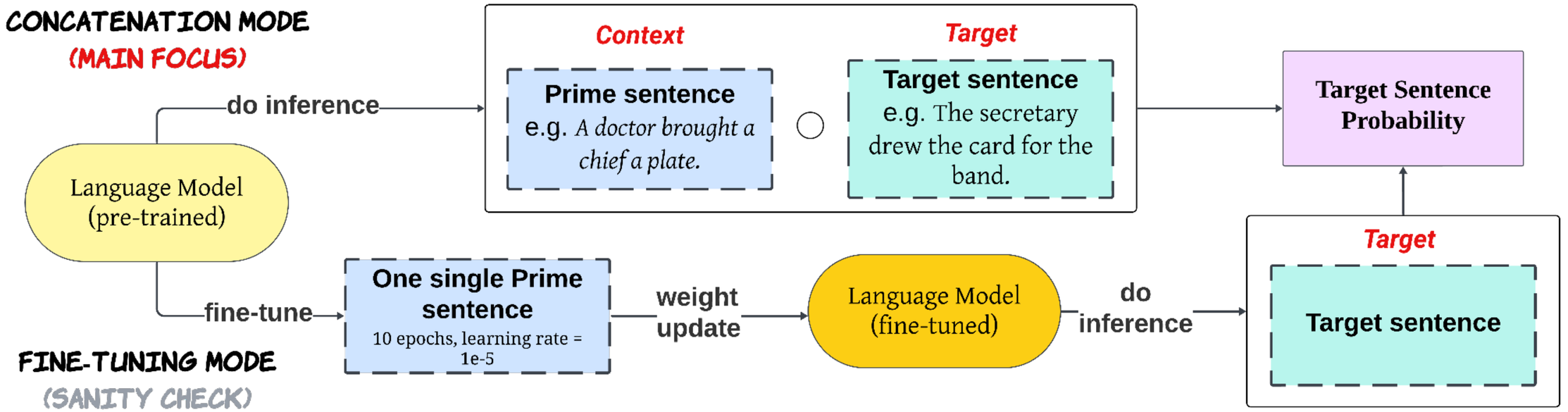
Experiment 1: Behavioral Analysis

Overview of the Current Approach



Fine-tuning Mode: [van Schijndal & Linzen 2018]
Concatenation Mode: [Sinclair et al. 2022]

Overview of the Current Approach



Assumption from Priming Theories: only some error-driven learning mechanism could lead to the IFE. Then, we predict that:

- Fine-tuning Mode: (with weight update) IFE ✅
- Concatenation Mode: (no weight update) IFE ❓

Fine-tuning Mode: [van Schijndal & Linzen 2018]
Concatenation Mode: [Sinclair et al. 2022]

Corpus

- 22 ditransitive verbs;
- 50 target sentences per verb;
- For each target sentence, pair it with a prime sentence with each prime verb;

22 x 50 (target sentences)

x

21 (prime sentences)

=

23100 <prime, target> pairs

Corpus

- 22 ditransitive verbs;
- 50 target sentences per verb;
- For each target sentence, pair it with a prime sentence with each prime verb;

22 x 50 (target sentences)

x

21 (prime sentences)

=

23100 <prime, target> pairs

Each <prime, target> pair ⇒ 4 structural combinations ⇒ 92400 trials.

$$\text{Prime}_{DO} \rightarrow \text{Target}_{DO}, \text{Prime}_{DO} \rightarrow \text{Target}_{PD}, \text{Prime}_{PD} \rightarrow \text{Target}_{DO}, \text{Prime}_{PD} \rightarrow \text{Target}_{PD}$$

Corpus

- 22 ditransitive verbs;
- 50 target sentences per verb;
- For each target sentence, pair it with a prime sentence with each prime verb;

22 x 50 (target sentences)

x

21 (prime sentences)

=

23100 <prime, target> pairs

Each <prime, target> pair ⇒ 4 structural combinations ⇒ 92400 trials.

$$\text{Prime}_{DO} \rightarrow \text{Target}_{DO}, \text{Prime}_{DO} \rightarrow \text{Target}_{PD}, \text{Prime}_{PD} \rightarrow \text{Target}_{DO}, \text{Prime}_{PD} \rightarrow \text{Target}_{PD}$$

Prime: A doctor brought a plate to a chief.

Target: The secretary drew the card for the band.

Quantifying Verb Biases and the IFE

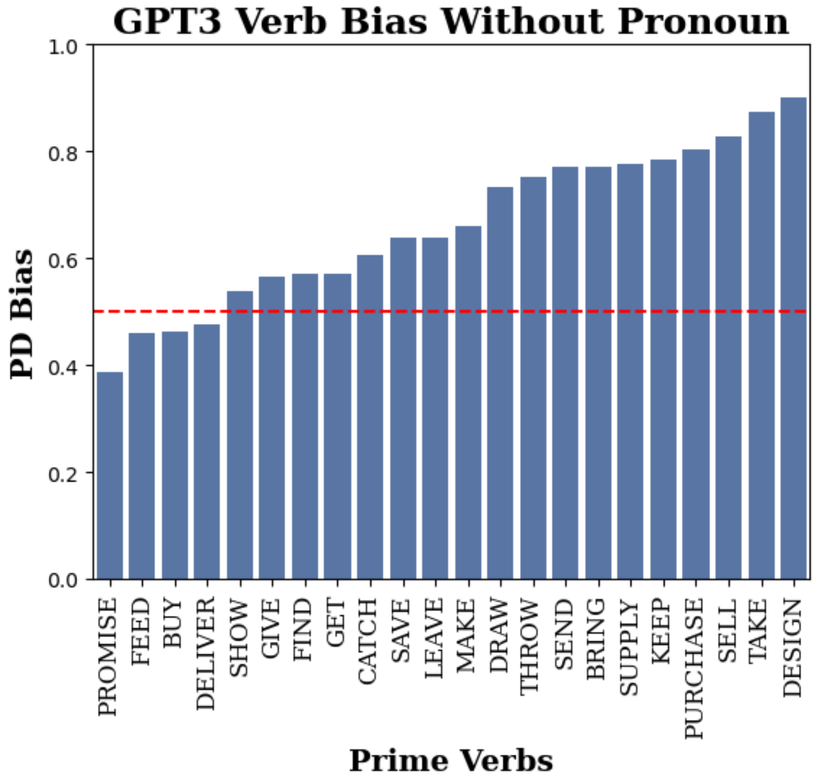
Verb Bias of verb V on structure X:

$$bias(V, \mathrm{PD}) = \frac{1}{|\mathcal{S}_V|} \sum_{t_{\mathrm{PD}} \in \mathcal{S}_V} \frac{\mathcal{P}(t_{\mathrm{PD}})}{\mathcal{P}(t_{\mathrm{PD}}) + \mathcal{P}(t_{\mathrm{DO}})}$$

Quantifying Verb Biases and the IFE

Verb Bias of verb V on structure X:

$$bias(V, \mathrm{PD}) = \frac{1}{|\mathcal{S}_V|} \sum_{t_{\mathrm{PD}} \in \mathcal{S}_V} \frac{\mathcal{P}(t_{\mathrm{PD}})}{\mathcal{P}(t_{\mathrm{PD}}) + \mathcal{P}(t_{\mathrm{DO}})}$$



Quantifying Verb Biases and the IFE

Verb Bias of verb V on structure X:

$$bias(V, \mathrm{PD}) = \frac{1}{|\mathcal{S}_V|} \sum_{t_{\mathrm{PD}} \in \mathcal{S}_V} \frac{\mathcal{P}(t_{\mathrm{PD}})}{\mathcal{P}(t_{\mathrm{PD}}) + \mathcal{P}(t_{\mathrm{DO}})}$$

Quantifying Verb Biases and the IFE

Verb Bias of verb V on structure X:

$$bias(V, \mathrm{PD}) = \frac{1}{|\mathcal{S}_V|} \sum_{t_{\mathrm{PD}} \in \mathcal{S}_V} \frac{\mathcal{P}(t_{\mathrm{PD}})}{\mathcal{P}(t_{\mathrm{PD}}) + \mathcal{P}(t_{\mathrm{DO}})}$$

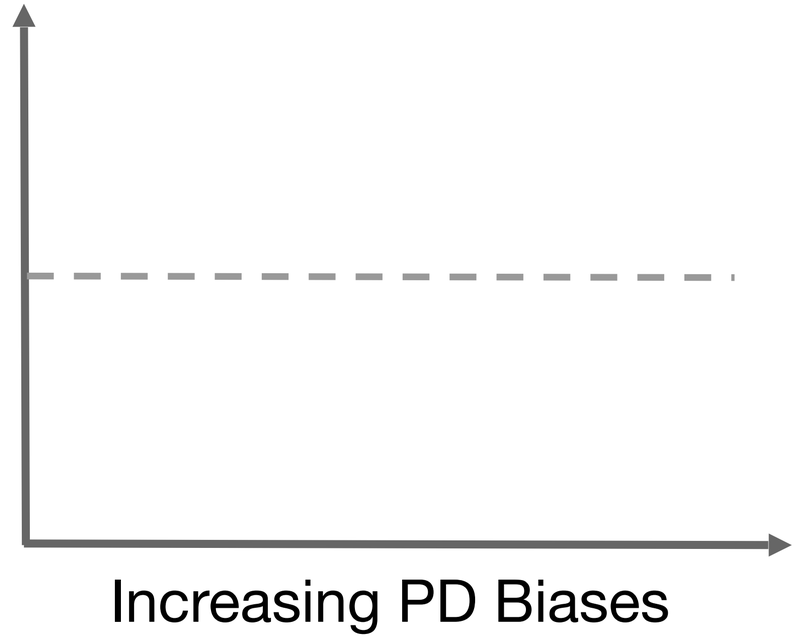
IFE: the priming effect for verb V in DO form on PD targets

$$PrimeBias(\mathrm{PD}|\mathrm{DO}, V) = \frac{1}{|T_{\mathrm{PD}}| \cdot |P^V_{\mathrm{DO}}|} \sum_{t_{\mathrm{PD}} \in T_{\mathrm{PD}}} \sum_{p^V_{\mathrm{DO}} \in P^V_{\mathrm{DO}}} \frac{\mathcal{P}(t_{\mathrm{PD}}|p^V_{\mathrm{DO}})}{\mathcal{P}(t_{\mathrm{DO}}|p^V_{\mathrm{DO}}) + \mathcal{P}(t_{\mathrm{PD}}|p^V_{\mathrm{DO}})}$$

Predictions on the IFE

Prime Verb	Verb PD Bias	Primed probability (priming magnitude)
Bring	0.23	0.23 (0.03)
Buy	0.27	0.25 (0.05)
Find	0.41	0.28 (0.08)
Draw	0.52	0.30 (0.10)
Design	0.77	0.33 (0.13)
priming in DO	*larger PD biases*	*greater priming effects*

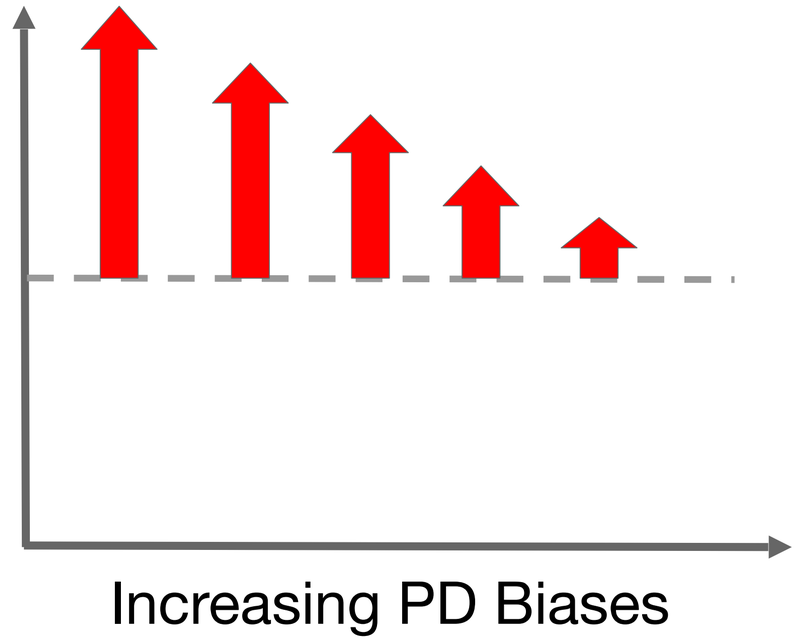
$\text{Prime}_{PD} \rightarrow \text{Target}_{PD}$



Predictions on the IFE

Prime Verb	Verb PD Bias	Primed probability (priming magnitude)
Bring	0.23	0.23 (0.03)
Buy	0.27	0.25 (0.05)
Find	0.41	0.28 (0.08)
Draw	0.52	0.30 (0.10)
Design	0.77	0.33 (0.13)
priming in DO	*larger PD biases*	*greater priming effects*

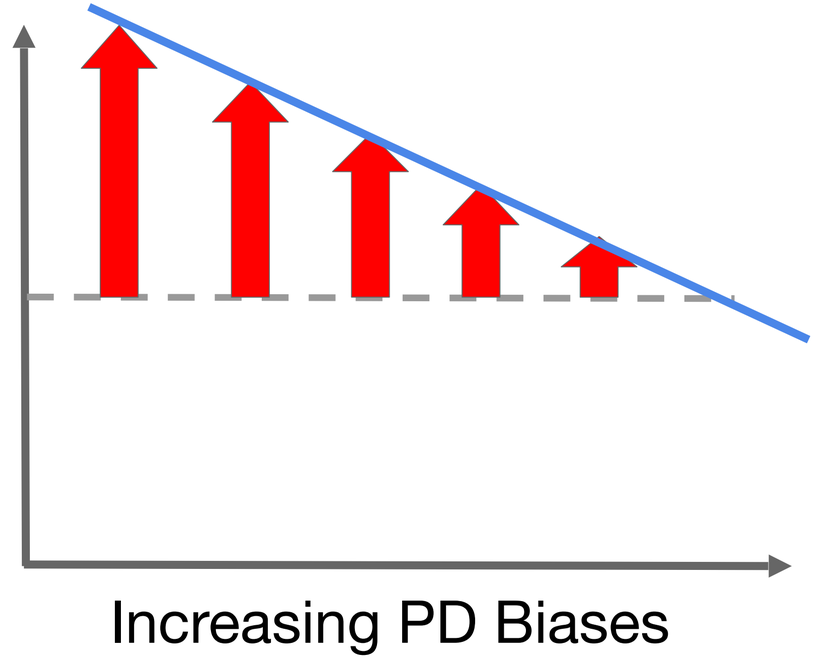
$\text{Prime}_{PD} \rightarrow \text{Target}_{PD}$



Predictions on the IFE

Prime Verb	Verb PD Bias	Primed probability (priming magnitude)
Bring	0.23	0.23 (0.03)
Buy	0.27	0.25 (0.05)
Find	0.41	0.28 (0.08)
Draw	0.52	0.30 (0.10)
Design	0.77	0.33 (0.13)
priming in DO	*larger PD biases*	*greater priming effects*

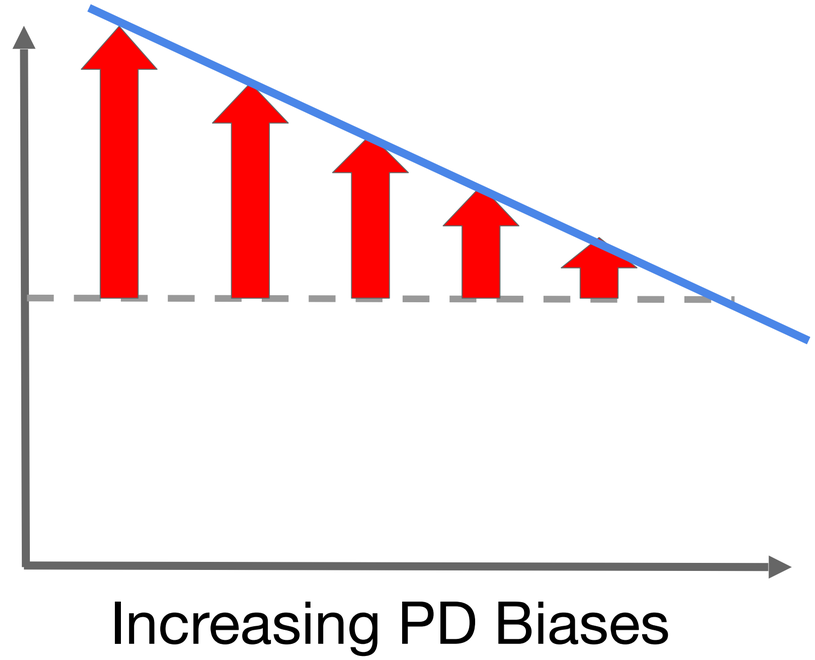
$\text{Prime}_{PD} \rightarrow \text{Target}_{PD}$



Predictions on the IFE

Prime Verb	Verb PD Bias	Primed probability (priming magnitude)
Bring	0.23	0.23 (0.03)
Buy	0.27	0.25 (0.05)
Find	0.41	0.28 (0.08)
Draw	0.52	0.30 (0.10)
Design	0.77	0.33 (0.13)
priming in DO	*larger PD biases*	*greater priming effects*

$\text{Prime}_{PD} \rightarrow \text{Target}_{PD}$



Increasing PD Biases

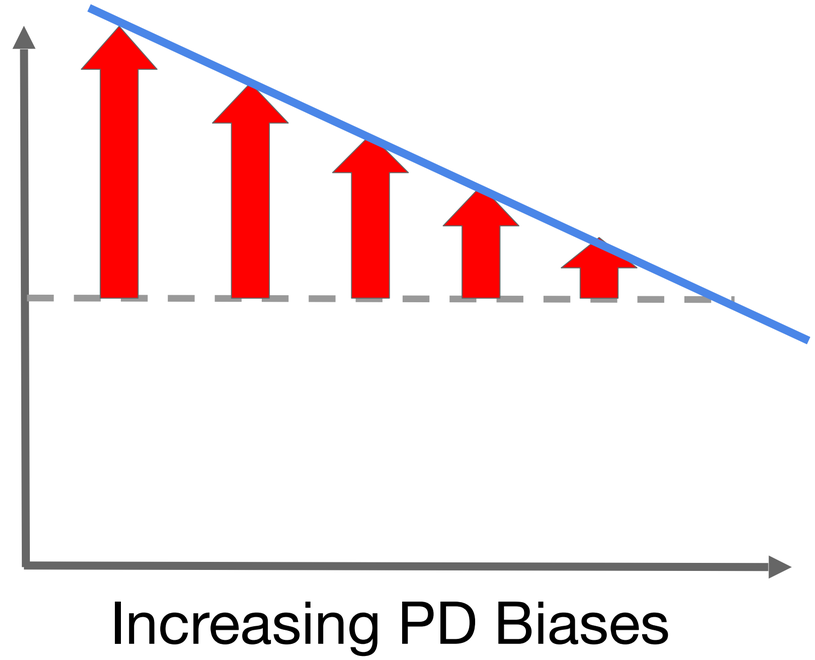
$\text{Prime}_{DO} \rightarrow \text{Target}_{PD}$

Increasing PD Biases

Predictions on the IFE

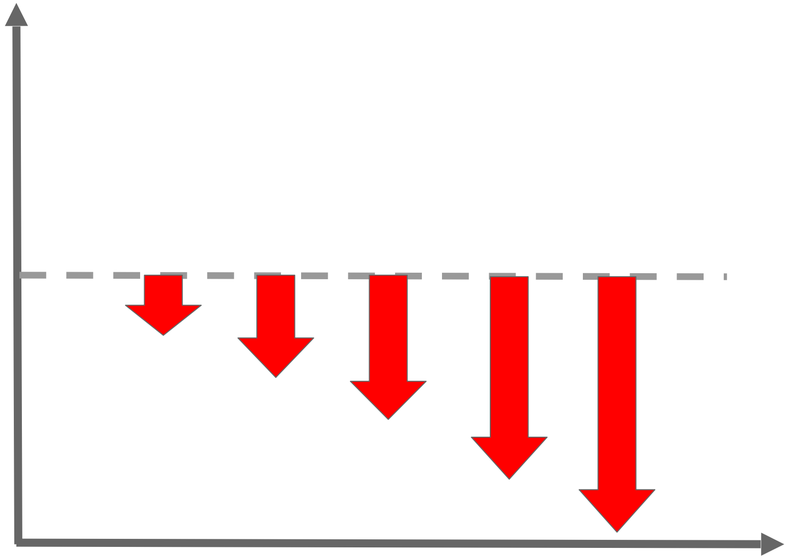
Prime Verb	Verb PD Bias	Primed probability (priming magnitude)
Bring	0.23	0.23 (0.03)
Buy	0.27	0.25 (0.05)
Find	0.41	0.28 (0.08)
Draw	0.52	0.30 (0.10)
Design	0.77	0.33 (0.13)
priming in DO	*larger PD biases*	*greater priming effects*

$\text{Prime}_{PD} \rightarrow \text{Target}_{PD}$



Increasing PD Biases

$\text{Prime}_{DO} \rightarrow \text{Target}_{PD}$

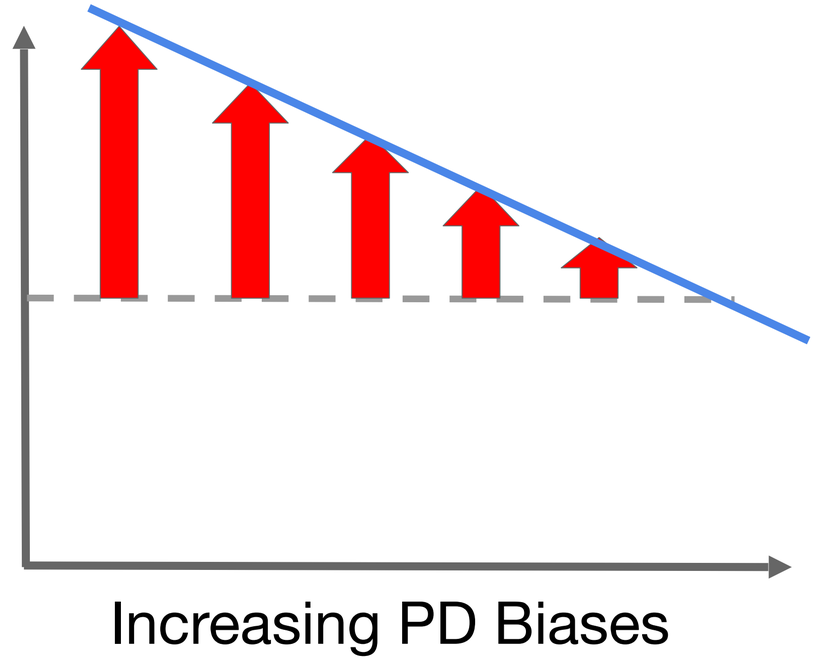


Increasing PD Biases

Predictions on the IFE

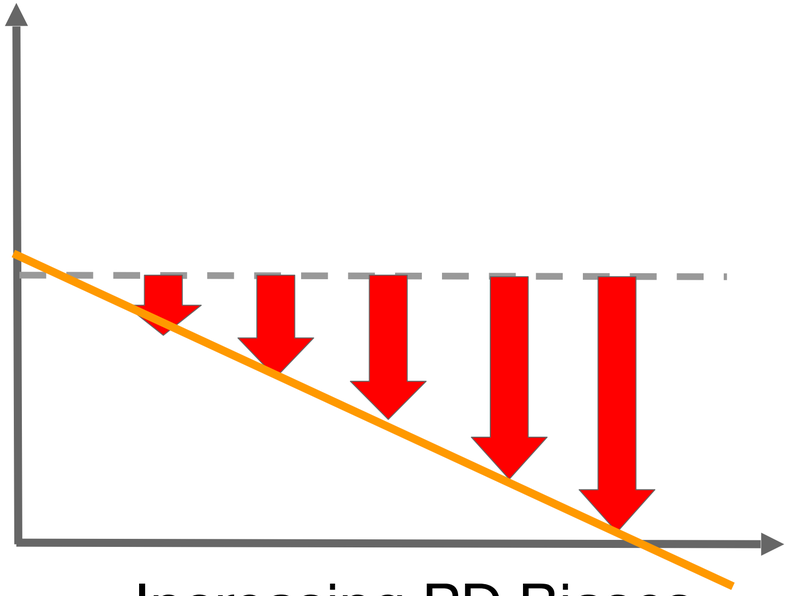
Prime Verb	Verb PD Bias	Primed probability (priming magnitude)
Bring	0.23	0.23 (0.03)
Buy	0.27	0.25 (0.05)
Find	0.41	0.28 (0.08)
Draw	0.52	0.30 (0.10)
Design	0.77	0.33 (0.13)
priming in DO	*larger PD biases*	*greater priming effects*

$\text{Prime}_{PD} \rightarrow \text{Target}_{PD}$



Increasing PD Biases

$\text{Prime}_{DO} \rightarrow \text{Target}_{PD}$

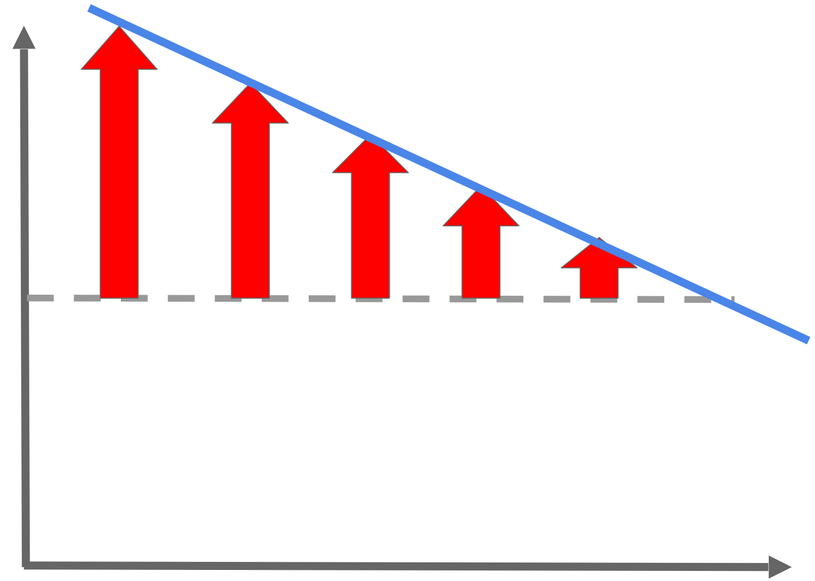


Increasing PD Biases

Predictions on the IFE

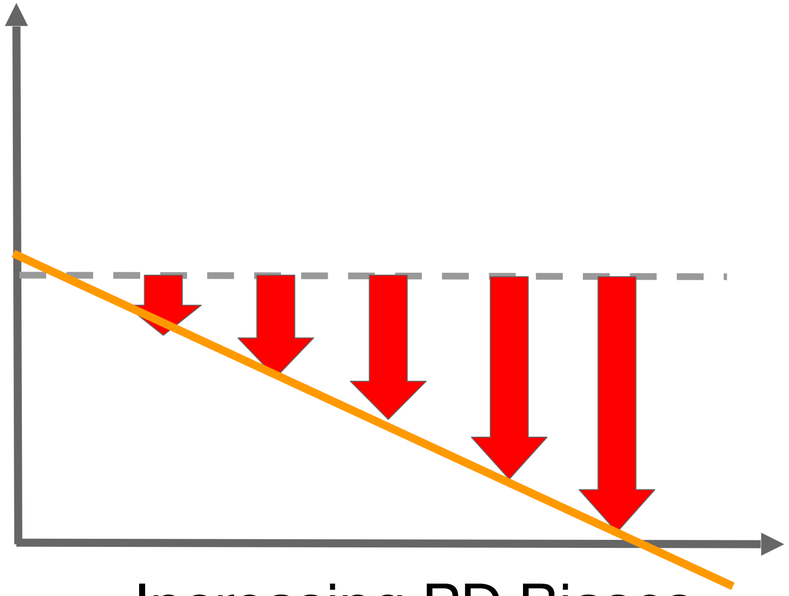
Prime Verb	Verb PD Bias	Primed probability (priming magnitude)
Bring	0.23	0.23 (0.03)
Buy	0.27	0.25 (0.05)
Find	0.41	0.28 (0.08)
Draw	0.52	0.30 (0.10)
Design	0.77	0.33 (0.13)
priming in DO	*larger PD biases*	*greater priming effects*

$\text{Prime}_{PD} \rightarrow \text{Target}_{PD}$



Increasing PD Biases

$\text{Prime}_{DO} \rightarrow \text{Target}_{PD}$



Increasing PD Biases

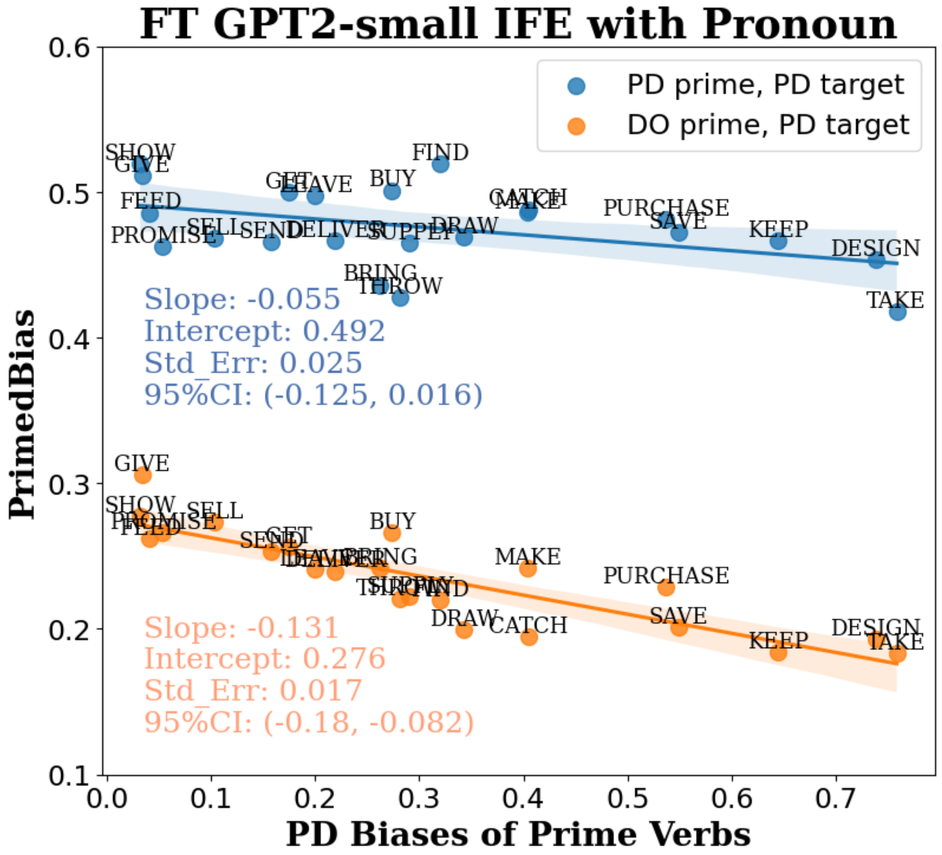
- **IFE**: double negative slopes;
- **Standard Priming**: PD-PD has higher intercept than DO-PD;

Priming in LLMs

Fine-tuning Mode: fine-tuning the model with the prime sentence and use the updated model to run the target sentence — with weight update;

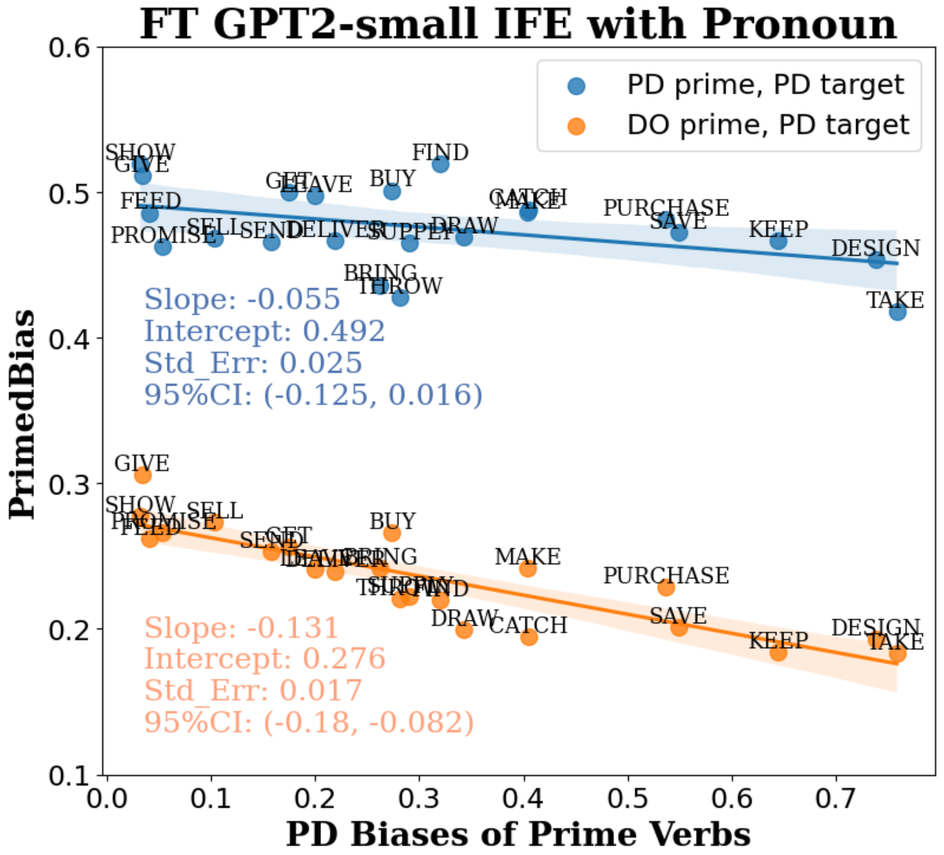
Priming in LLMs

Fine-tuning Mode: fine-tuning the model with the prime sentence and use the updated model to run the target sentence — with weight update;



Priming in LLMs

Fine-tuning Mode: fine-tuning the model with the prime sentence and use the updated model to run the target sentence — with weight update;



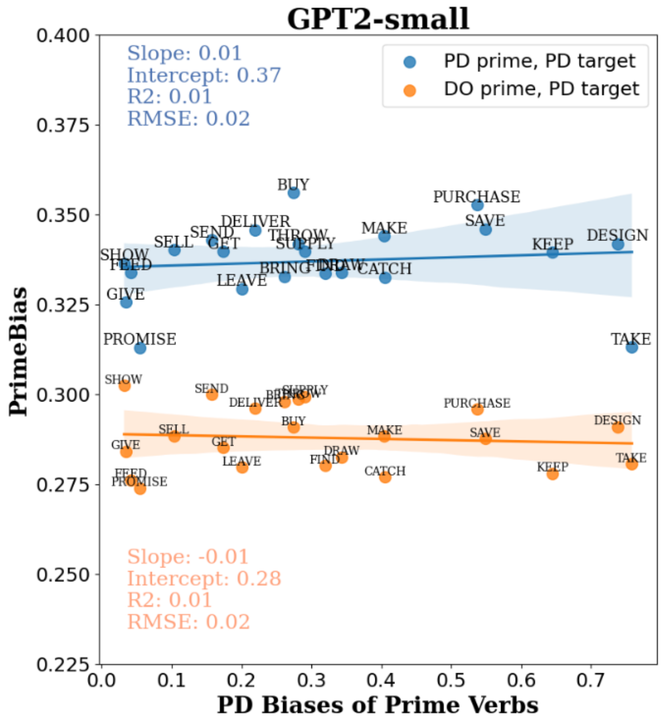
Even *GPT2-small* shows significant inverse frequency effects!

Priming in LLMs

Concatenation Mode: concatenating the prime and target sentences and run the model — without weight update;

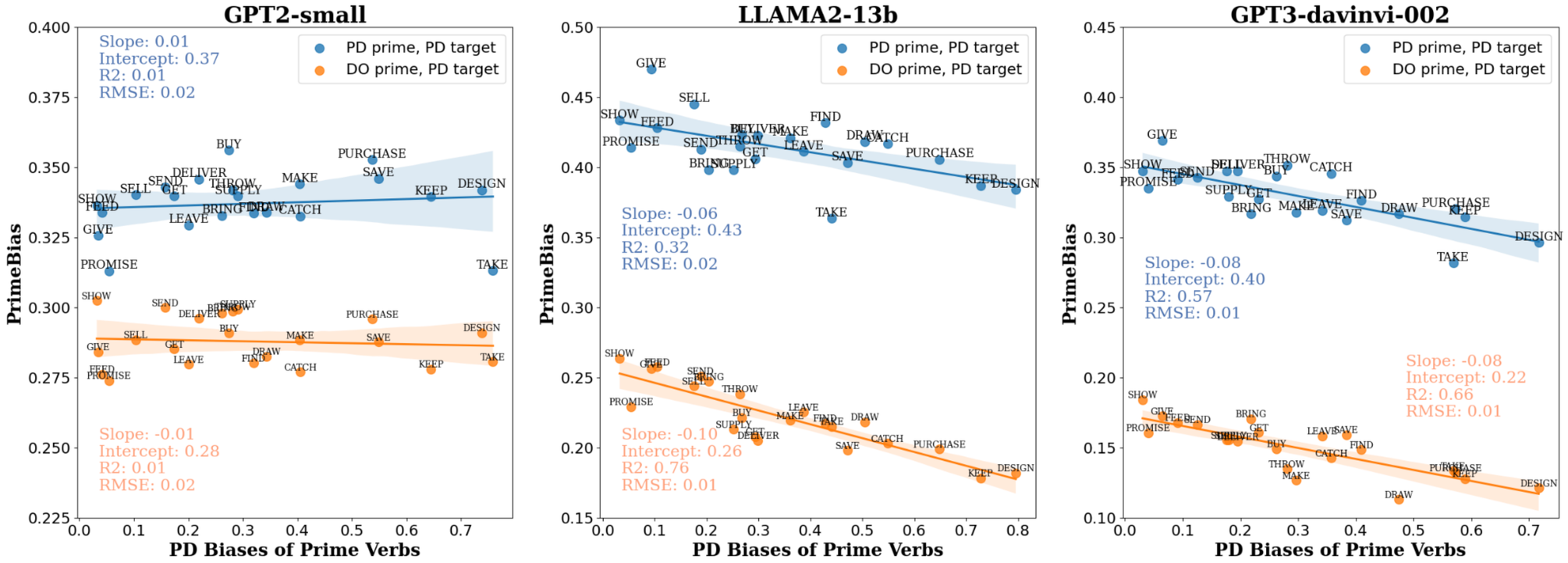
Priming in LLMs

Concatenation Mode: concatenating the prime and target sentences and run the model — without weight update;



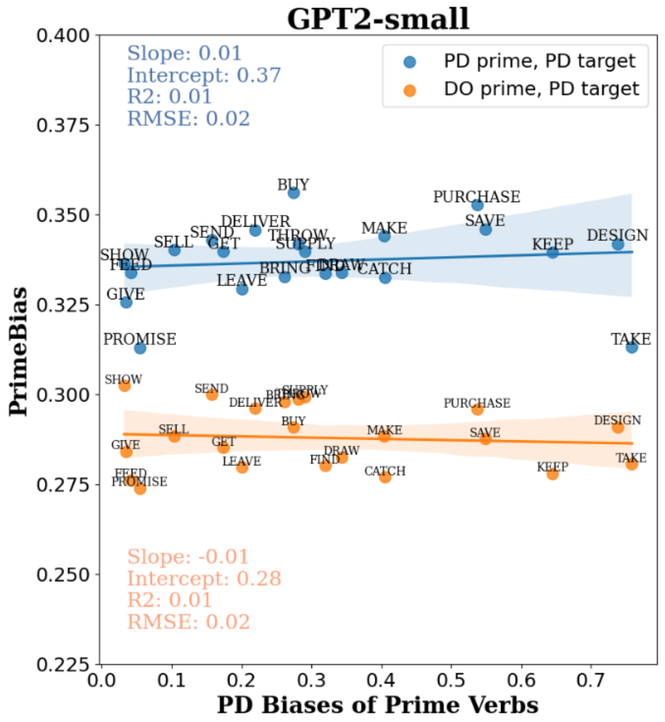
Priming in LLMs

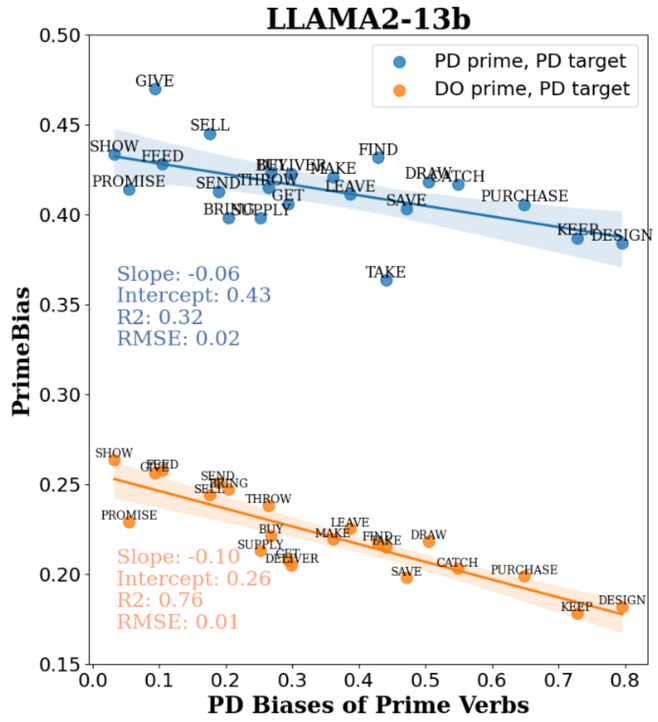
Concatenation Mode: concatenating the prime and target sentences and run the model — without weight update;

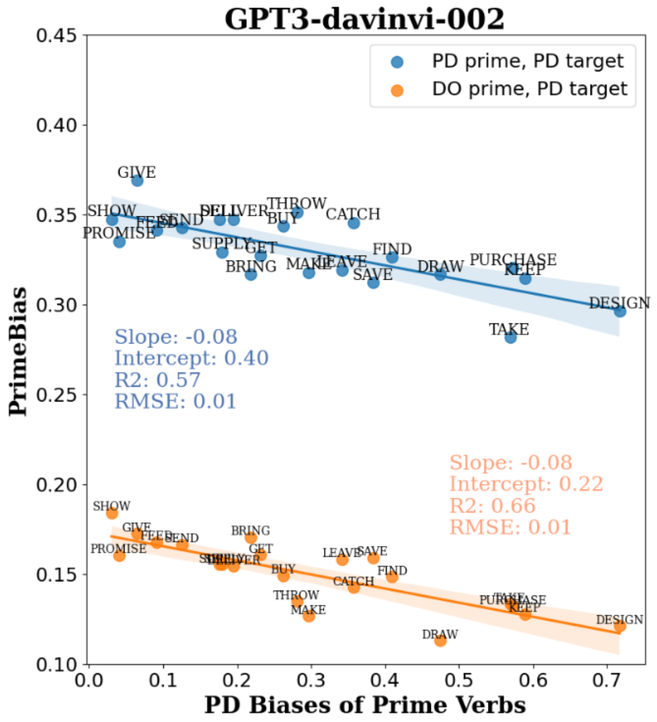


Priming in LLMs

Concatenation Mode: concatenating the prime and target sentences and run the model — without weight update;

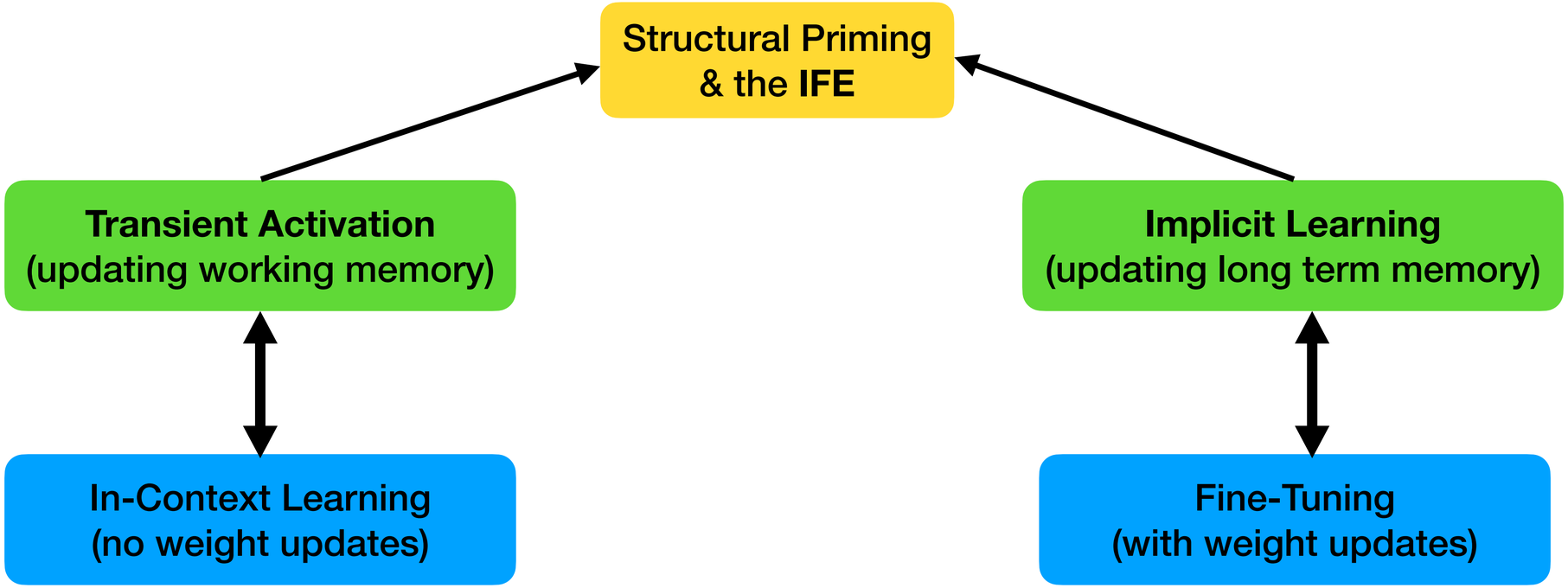




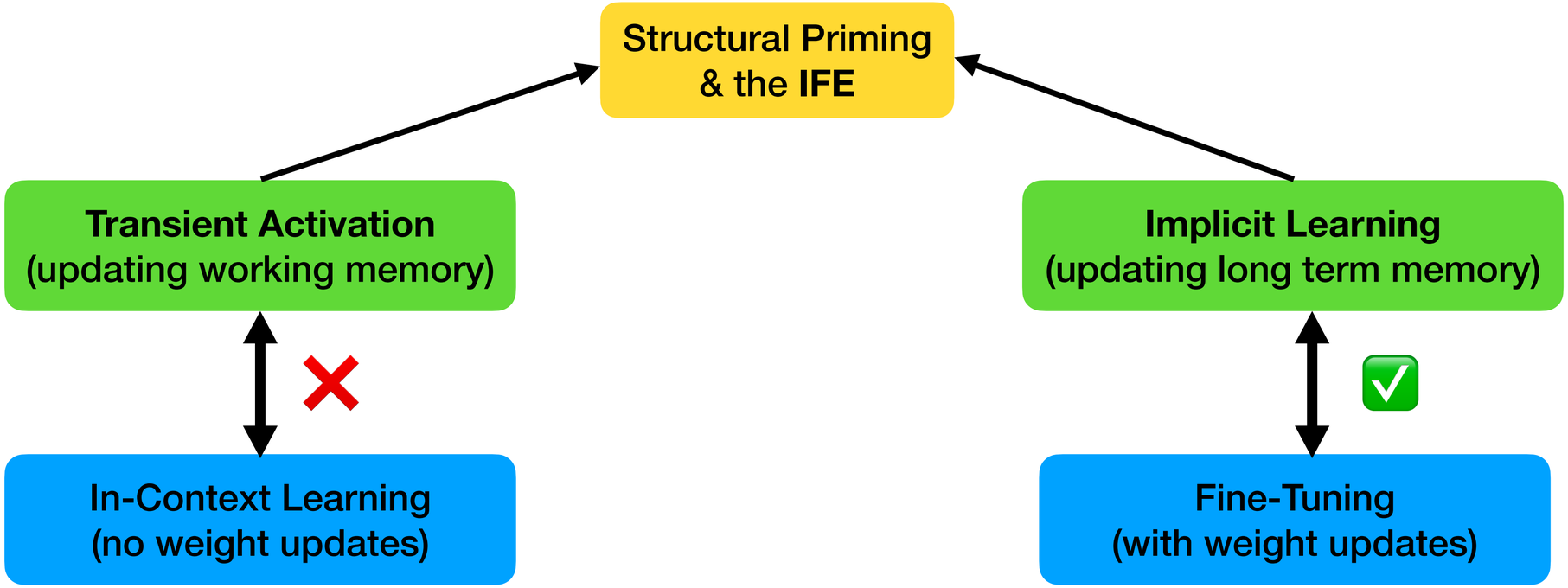


Larger models show more significant inverse frequency effects!

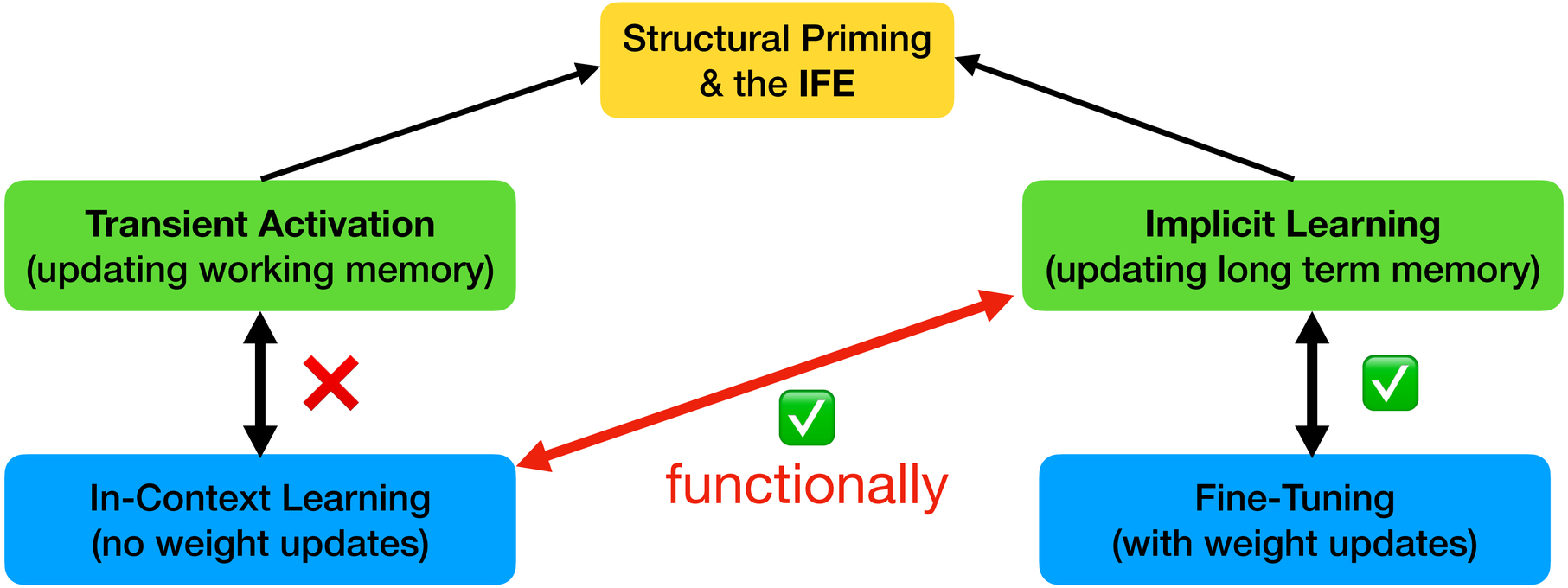
Takeaways from inverse frequency



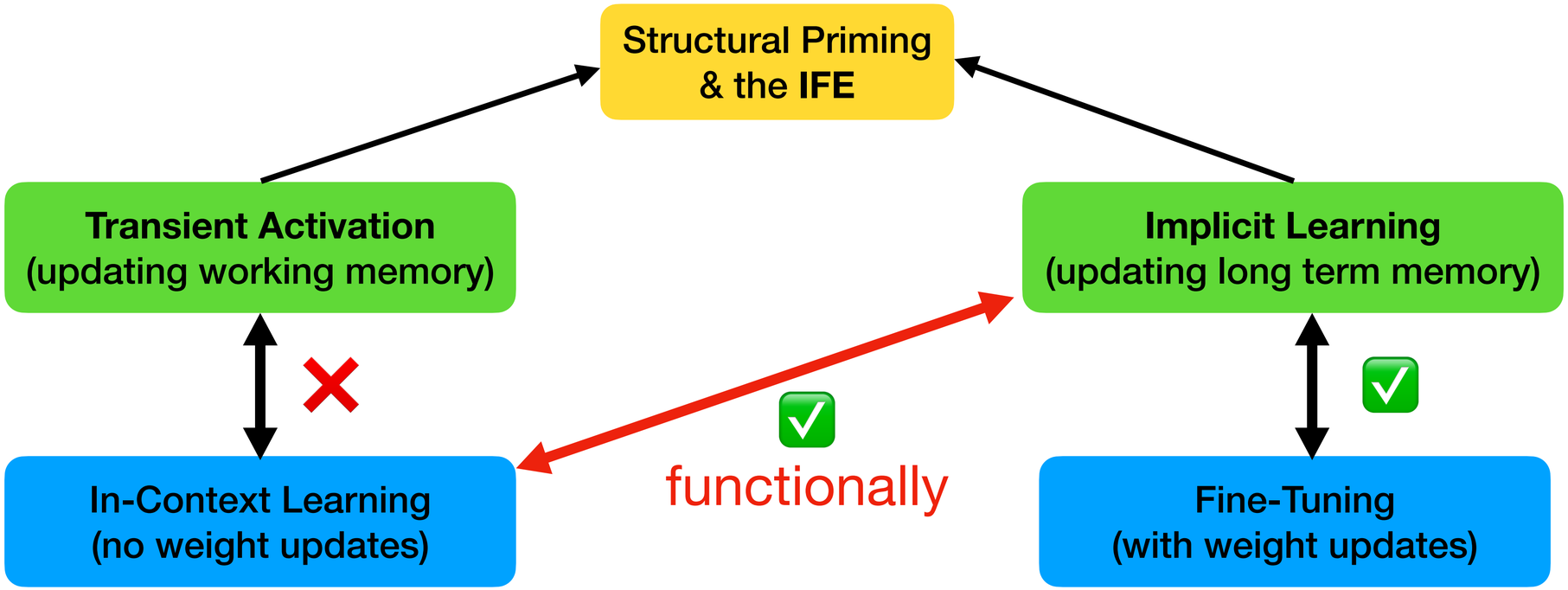
Takeaways from inverse frequency



Takeaways from inverse frequency



Takeaways from inverse frequency



We used the IFE as a diagnostic on the error-driven nature of ICL as a processing / adaptation mechanism in LLMs.

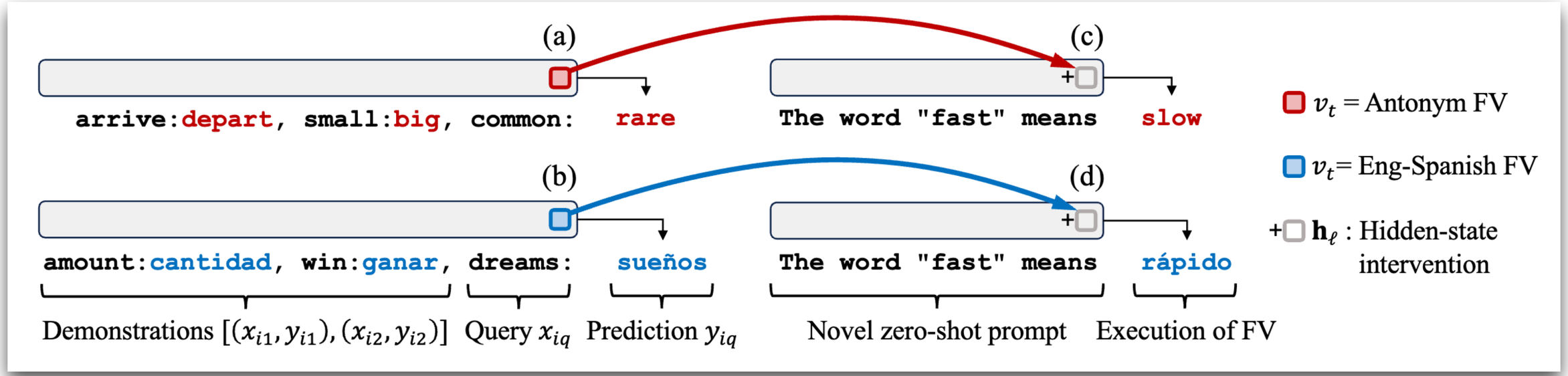
Experiment 2: Mechanistic Analysis

Research Questions:

- How to further characterize the generalized “ICL ≈ Priming” claim?
- How could we causally verify the involvement of verb bias?

ICL leads to Function Vectors

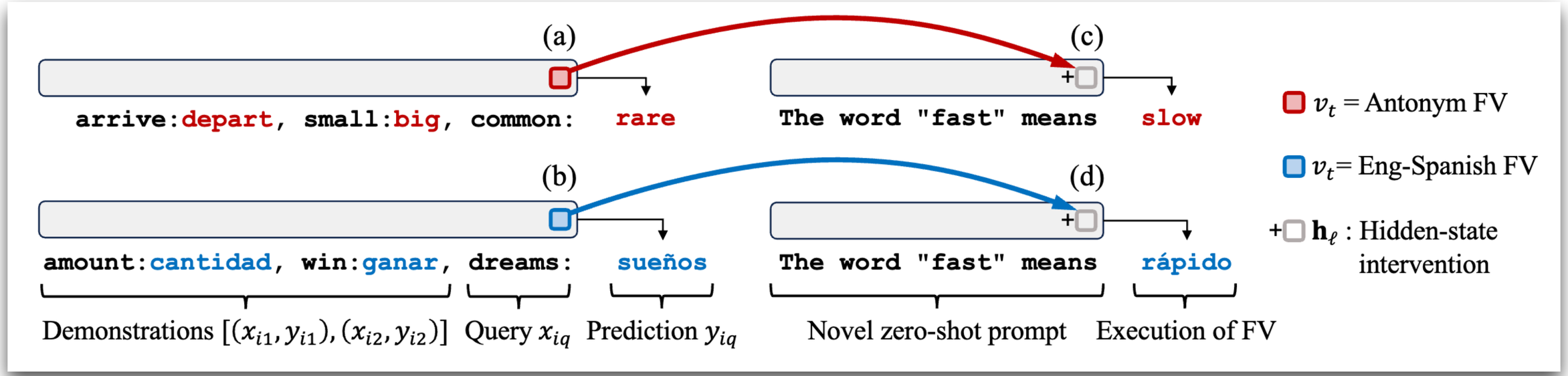
Function vectors: compressed vector representation of the task in context;



[e.g., Todd et al. 2024; Hendel et al. 2023]

ICL leads to Function Vectors

Function vectors: compressed vector representation of the task in context;

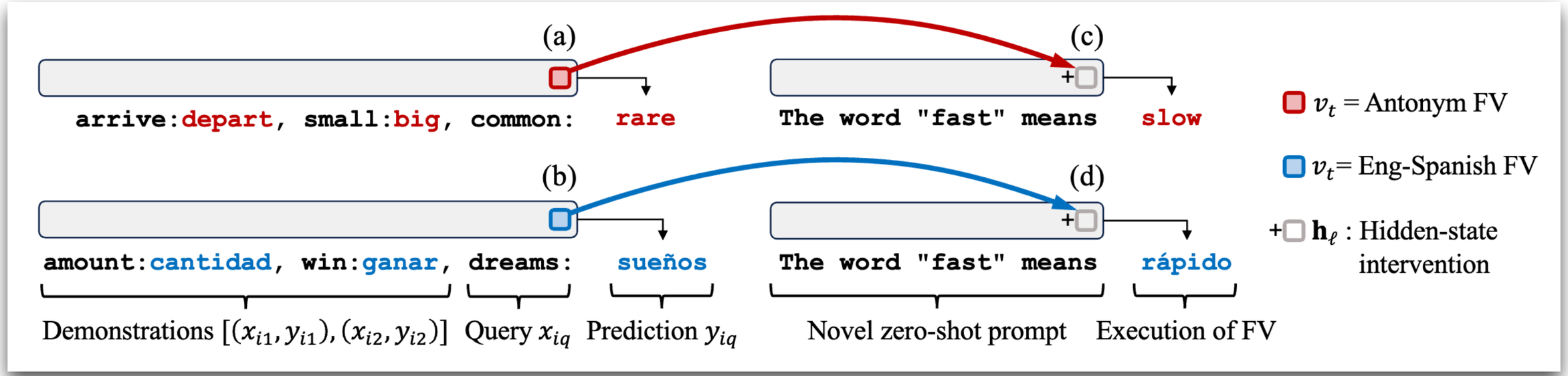


Extract mean activations at the query position after context

[e.g., Todd et al. 2024; Hendel et al. 2023]

ICL leads to Function Vectors

Function vectors: compressed vector representation of the task in context;

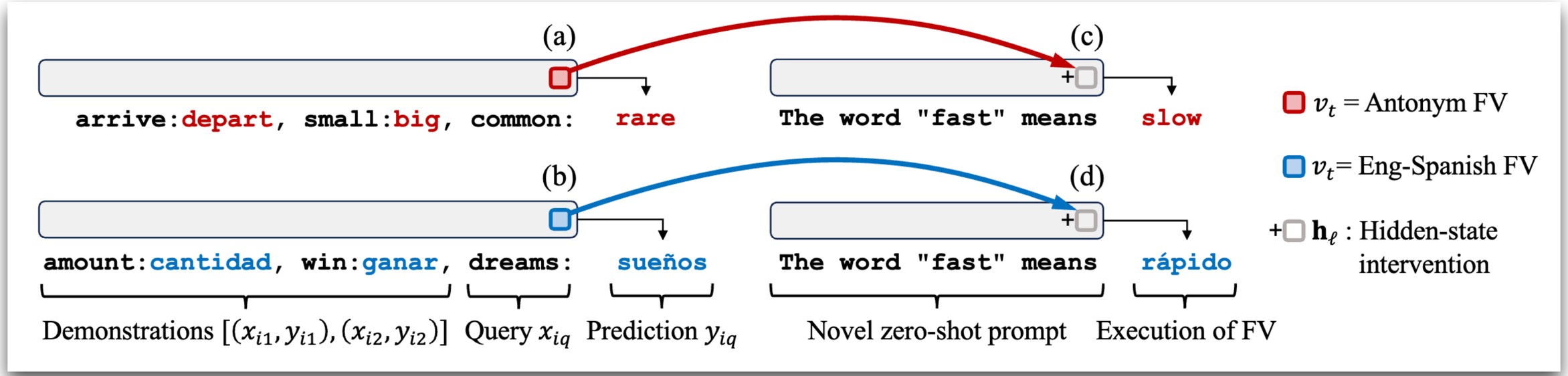


Extract mean activations at the query position after context → Injecting this vector in a new inference run (zero-shot)

[e.g., Todd et al. 2024; Hendel et al. 2023]

ICL leads to Function Vectors

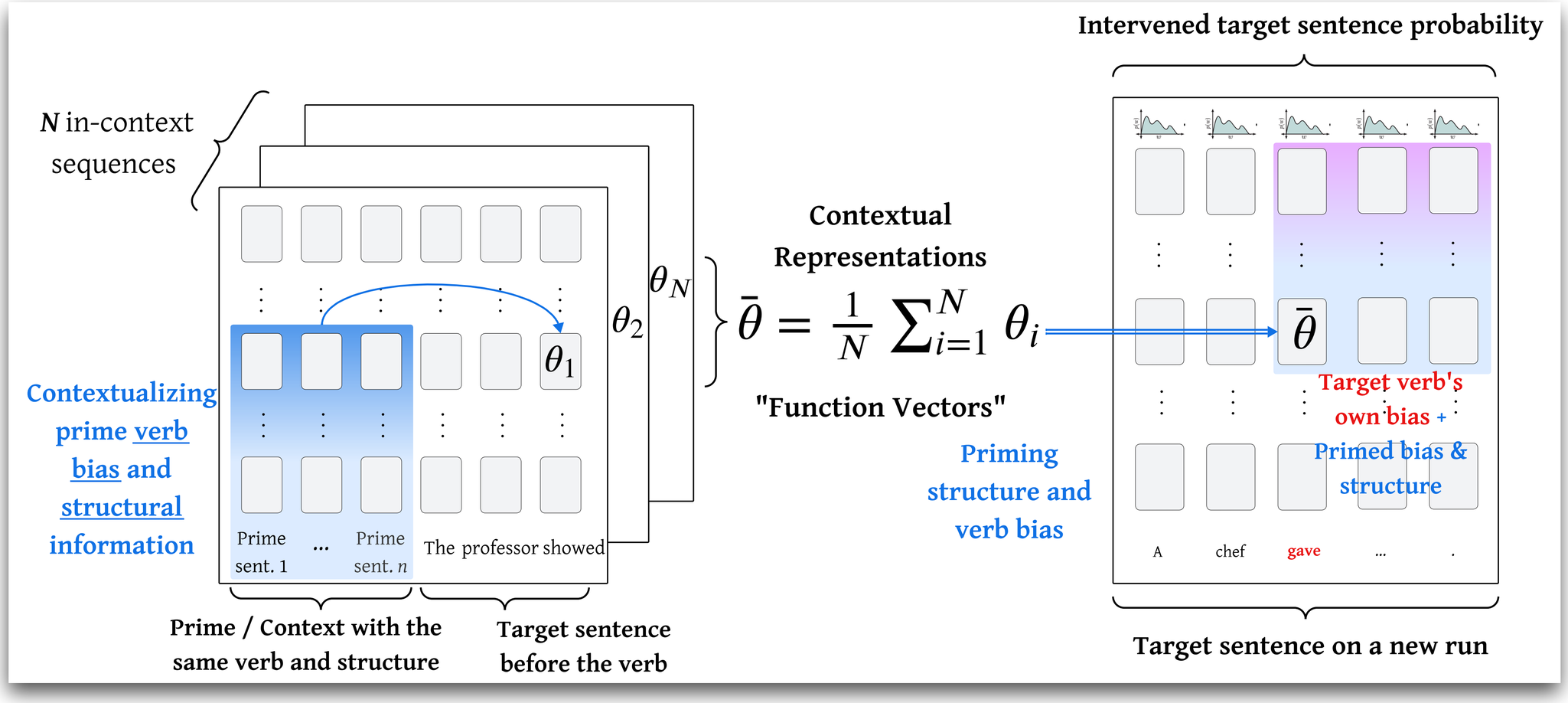
Function vectors: compressed vector representation of the task in context;



Extract mean activations at the query position after context → Injecting this vector in a new inference run (zero-shot) → Steering model production towards the task

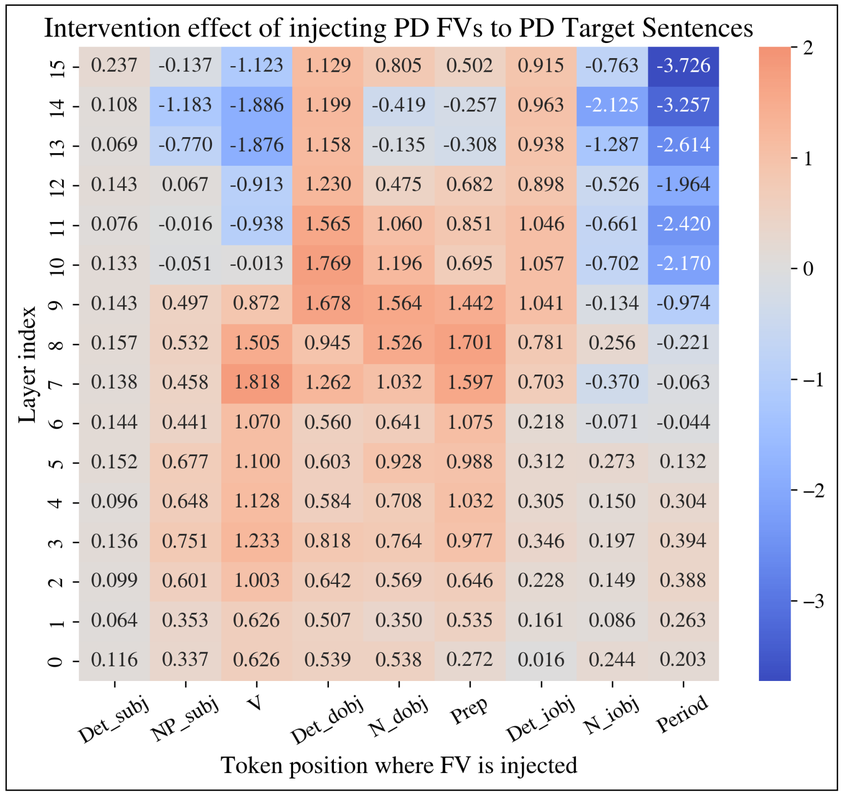
[e.g., Todd et al. 2024; Hendel et al. 2023]

Structural Priming with Function Vectors

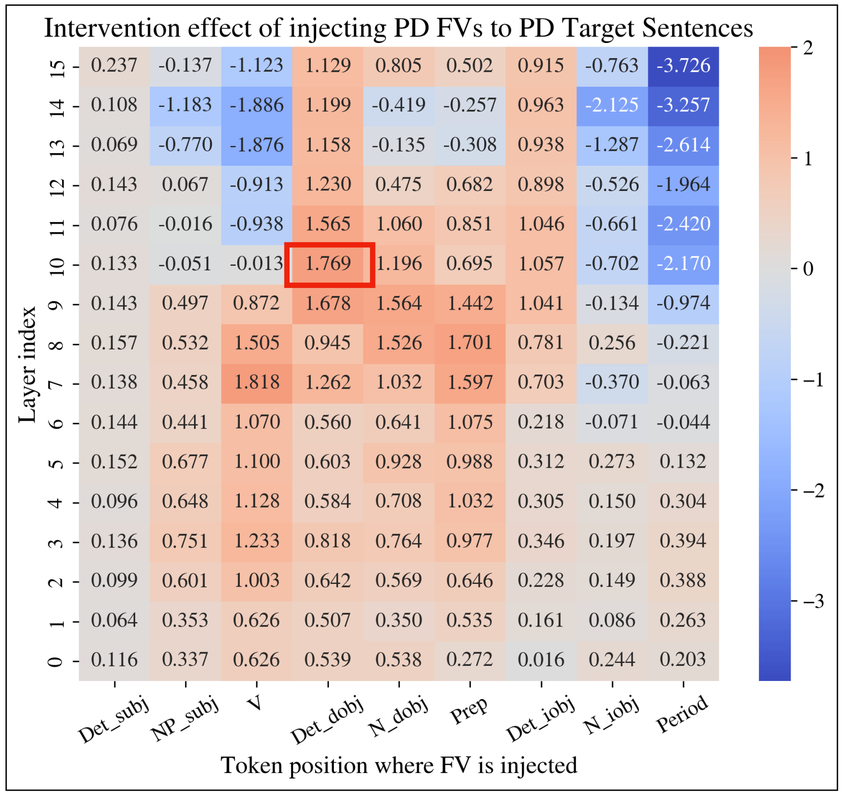


* we also did a version where function vectors are the mean of the most influential attention heads' output activations (Todd et al. 2024), leading to similar results.

Structural Priming with Function Vectors

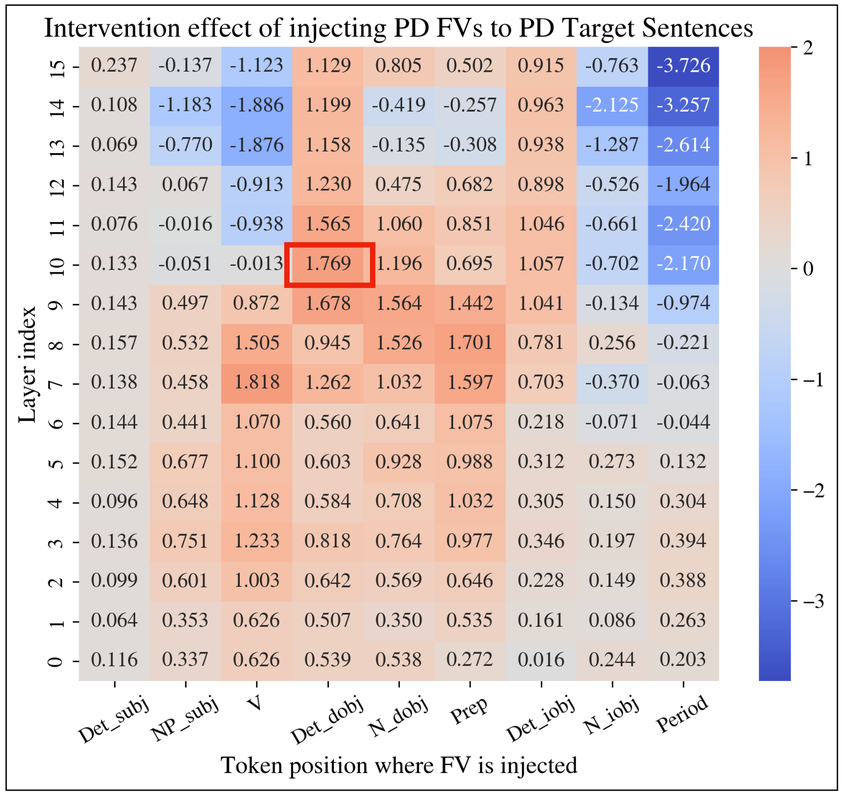


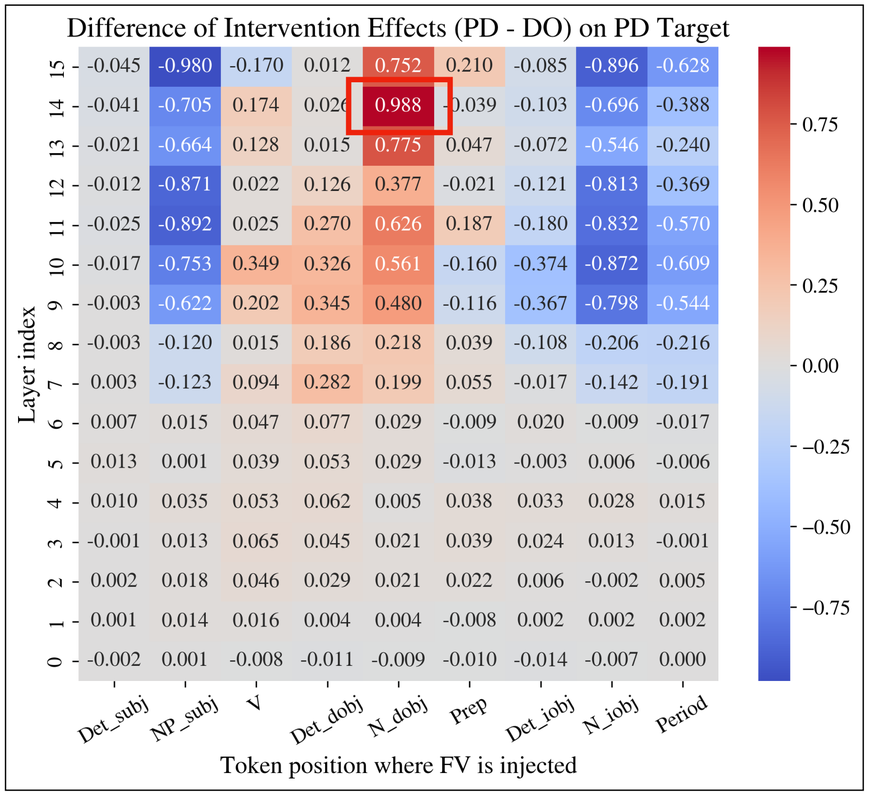
Structural Priming with Function Vectors



- Observing the priming effect: FVs increased target sentence probability;

Structural Priming with Function Vectors

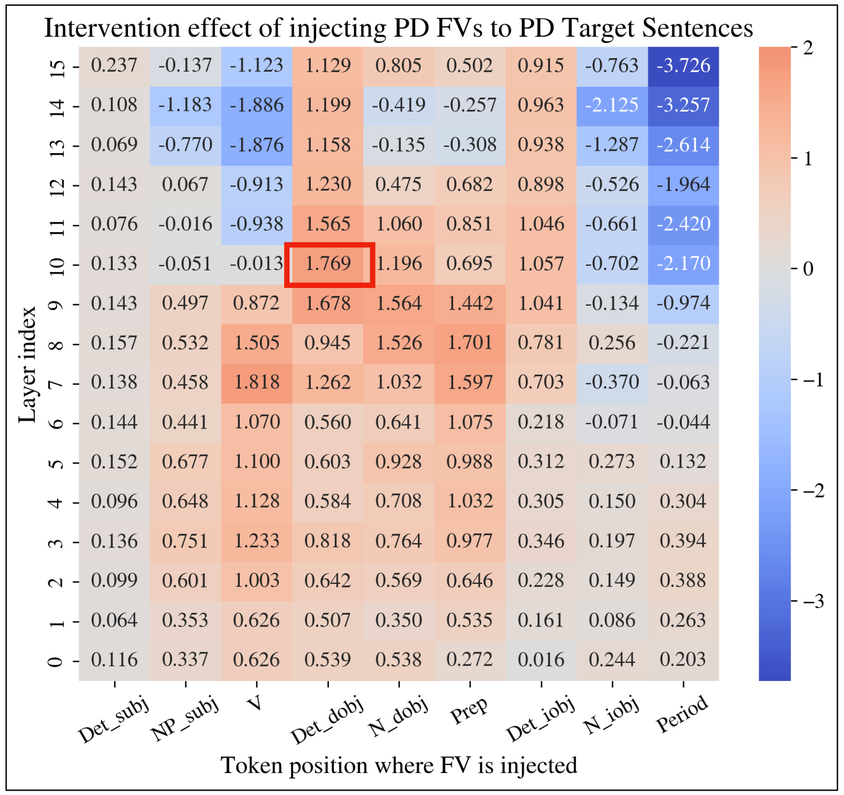


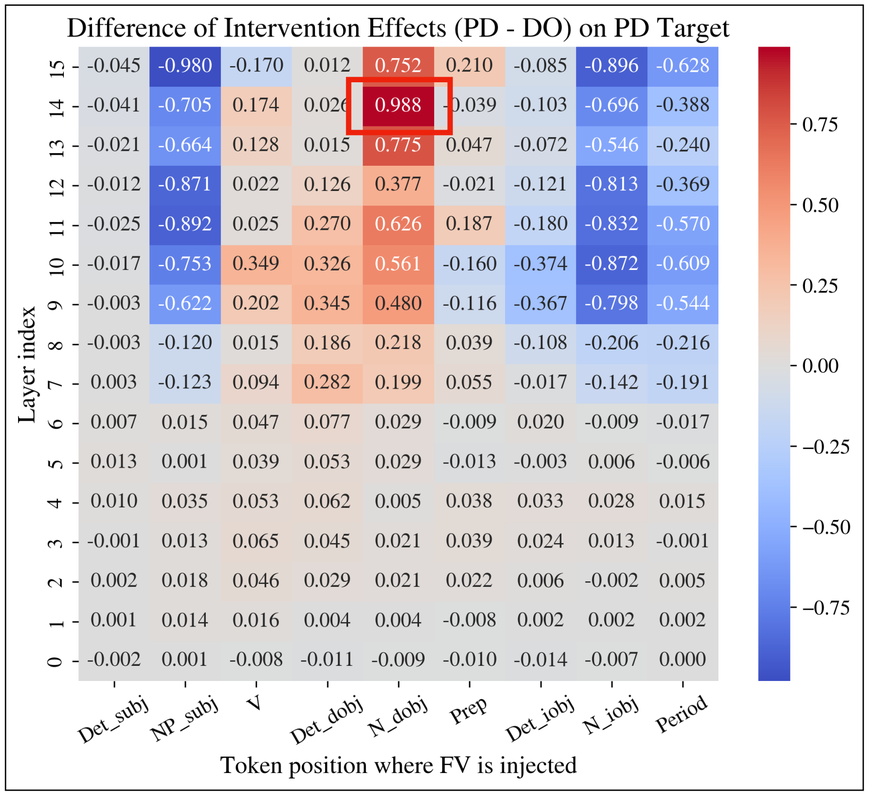


- Observing the priming effect: FVs increased target sentence probability;

- PD FVs cause a larger intervention effect than DO FVs on PD targets.

Structural Priming with Function Vectors

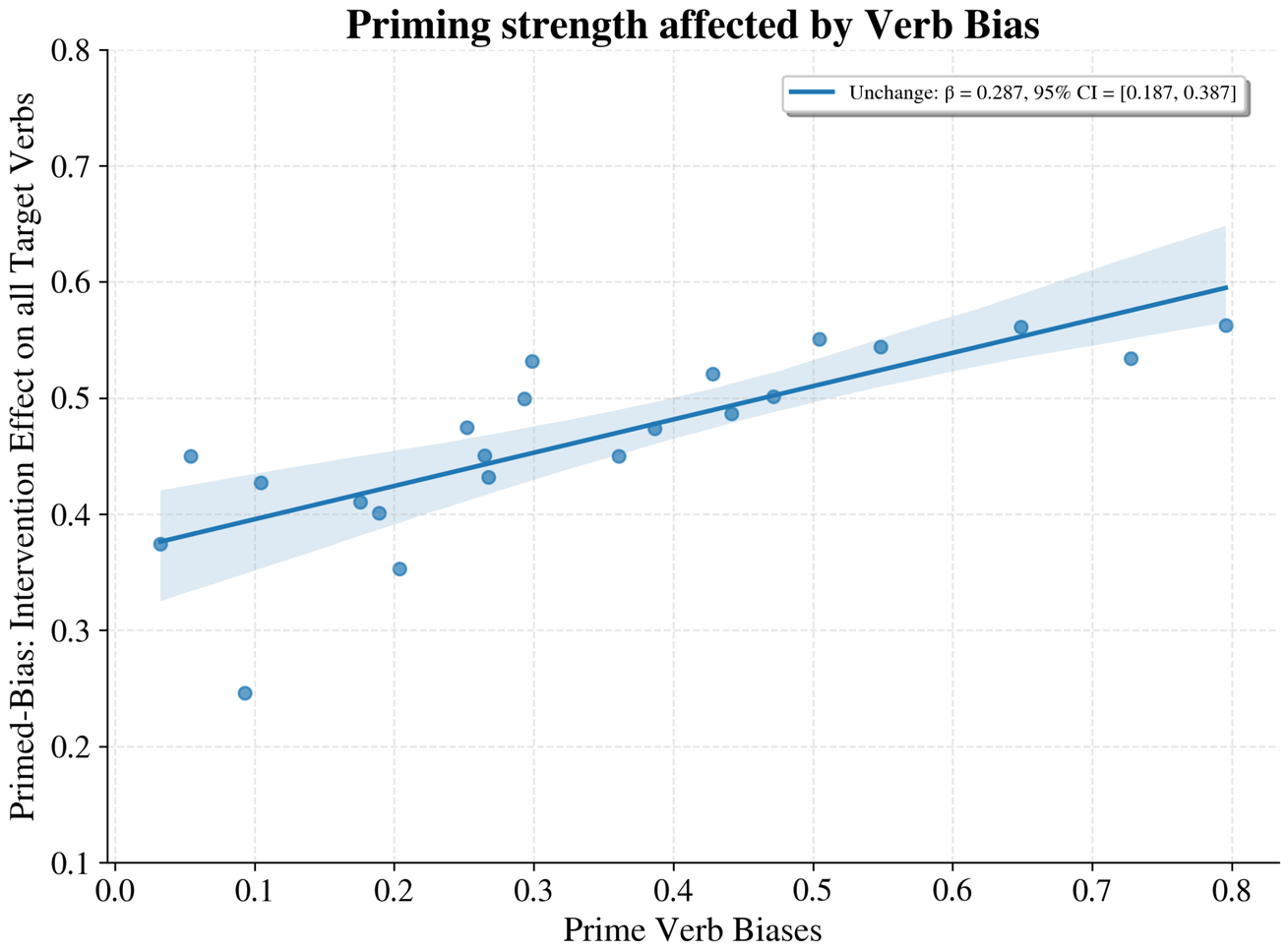




- Observing the priming effect: FVs increased target sentence probability;
- PD FVs cause a larger intervention effect than DO FVs on PD targets.

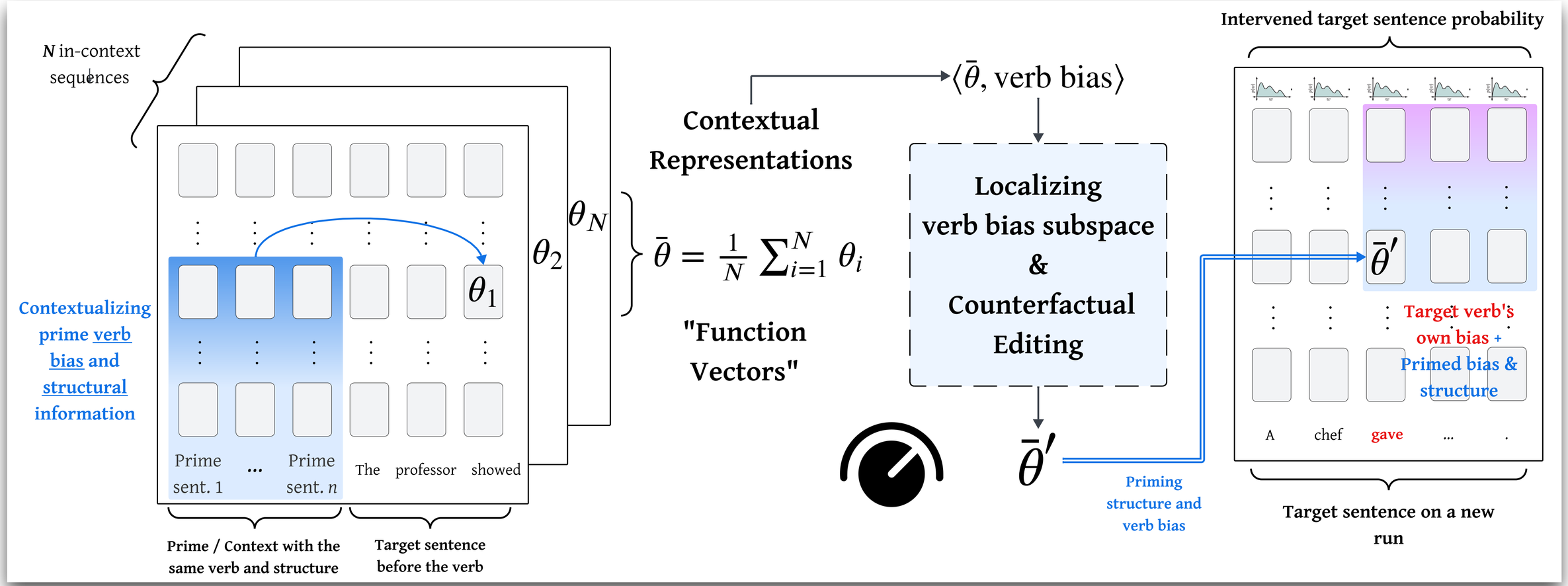
Yes, we do observe standard priming effect.

Structural Priming with Function Vectors

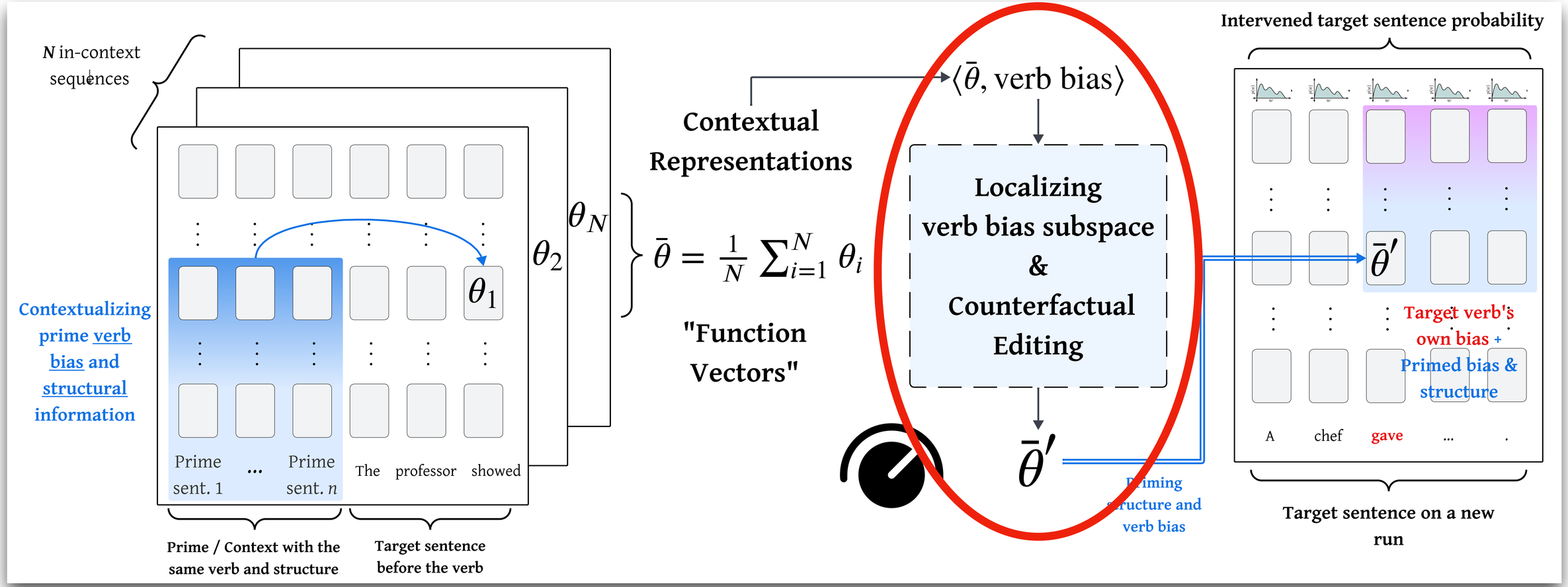


However, no IFE is observed…!

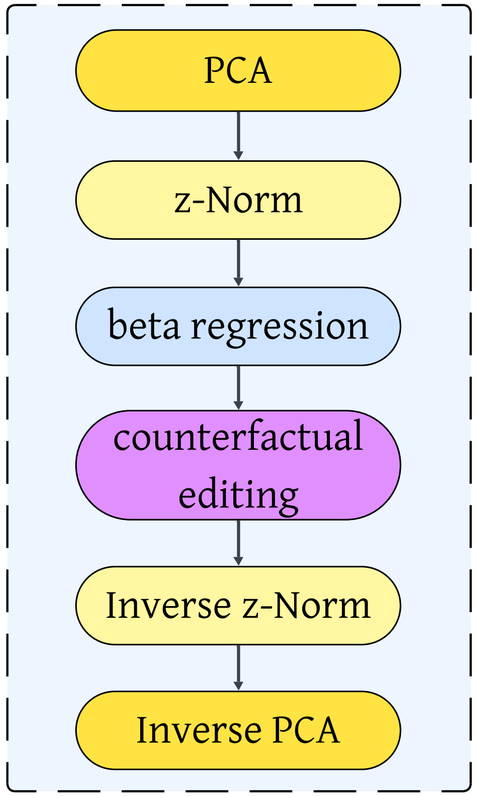
Causal Role of Verb Bias in Function Vectors?

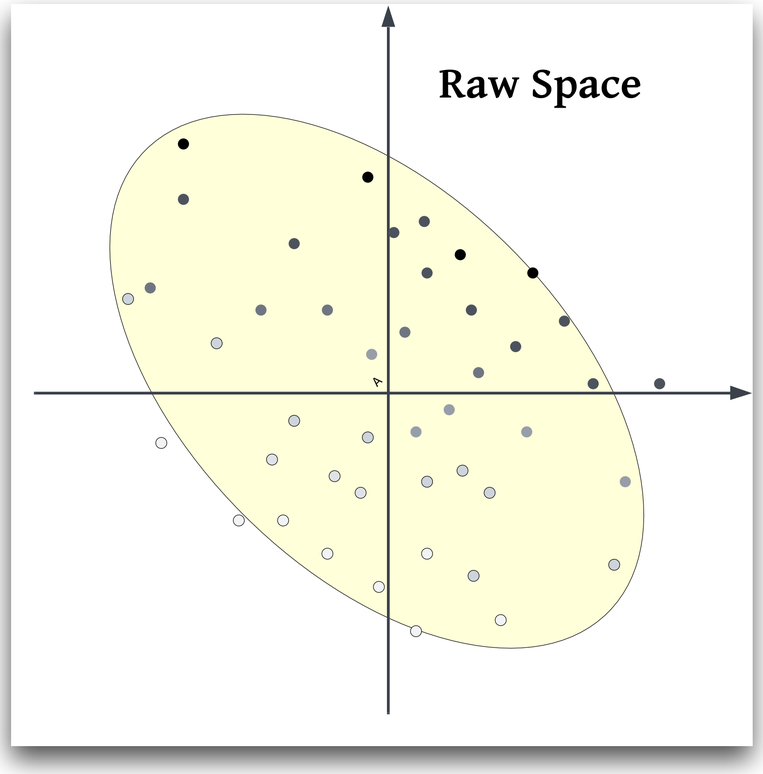


Causal Role of Verb Bias in Function Vectors?



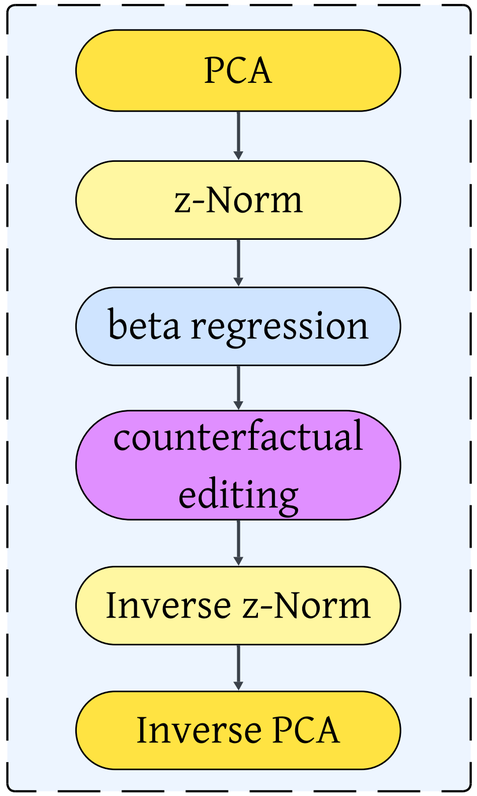
A Method for Continuous Variable Intervention

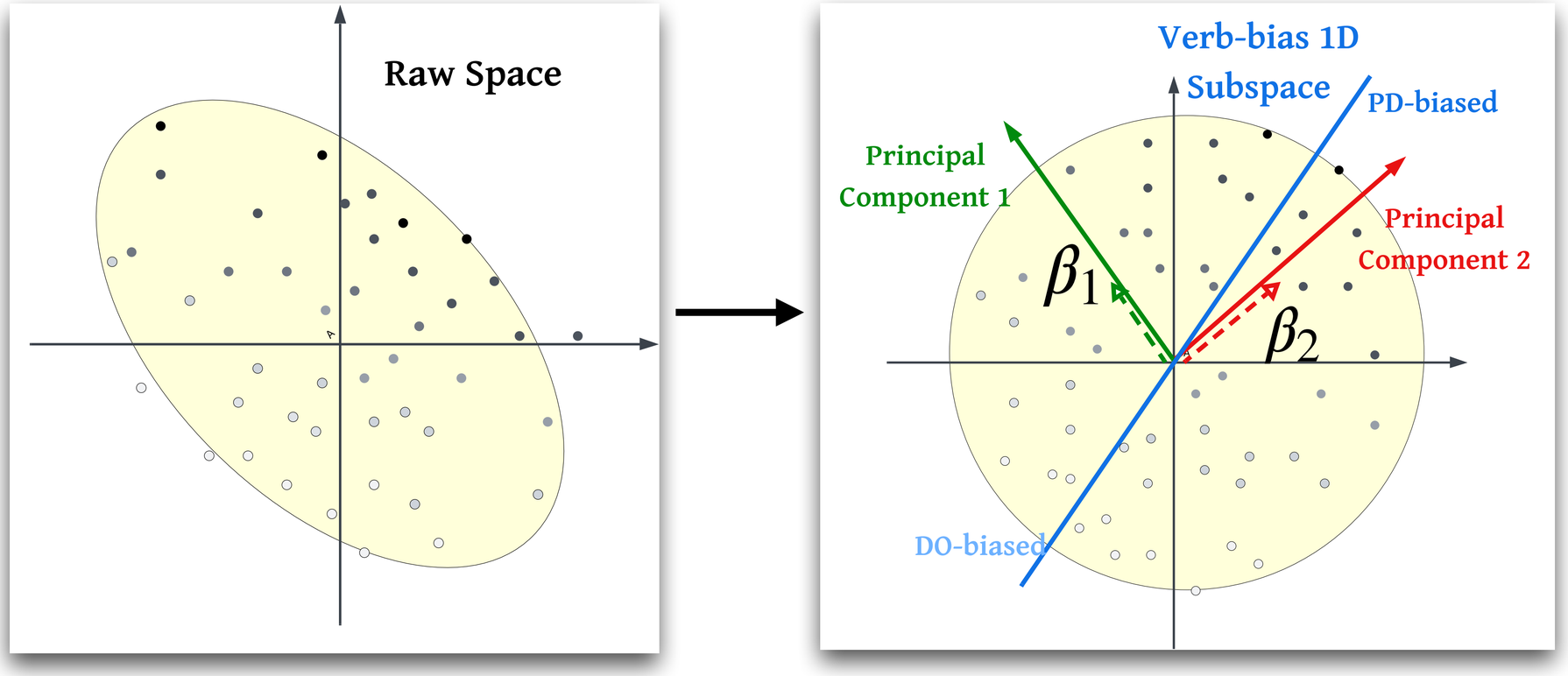




[For binary/discrete variable intervention, see Ravfogel et al. 2021, Hao & Linzen 2023, Ozaki et al. 2025]

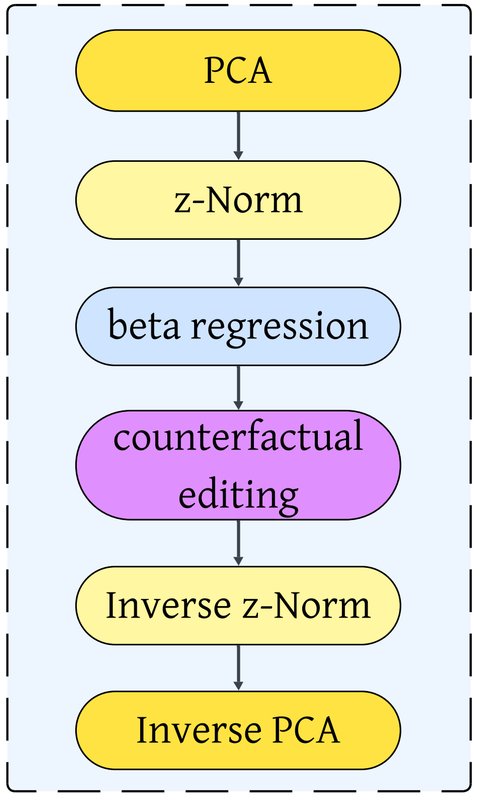
A Method for Continuous Variable Intervention

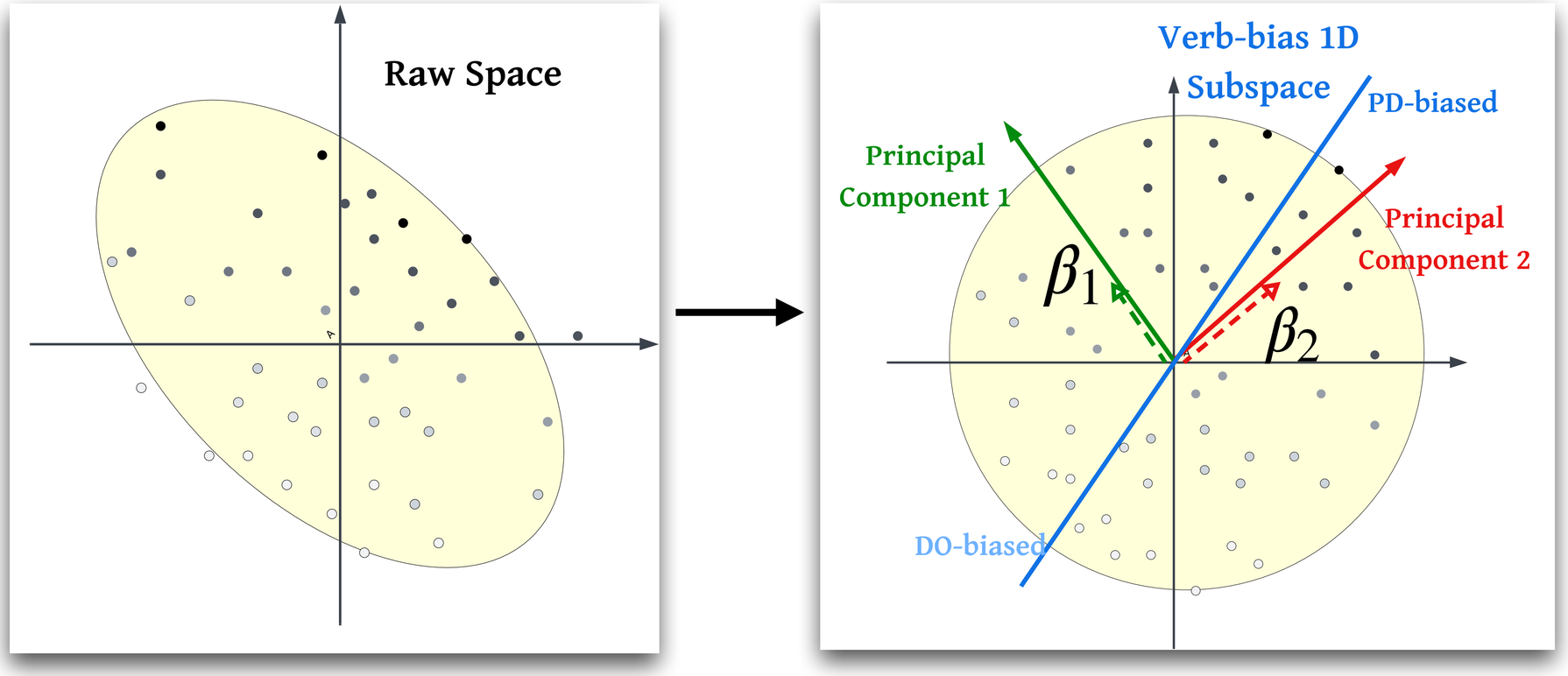




[For binary/discrete variable intervention, see Ravfogel et al. 2021, Hao & Linzen 2023, Ozaki et al. 2025]

A Method for Continuous Variable Intervention

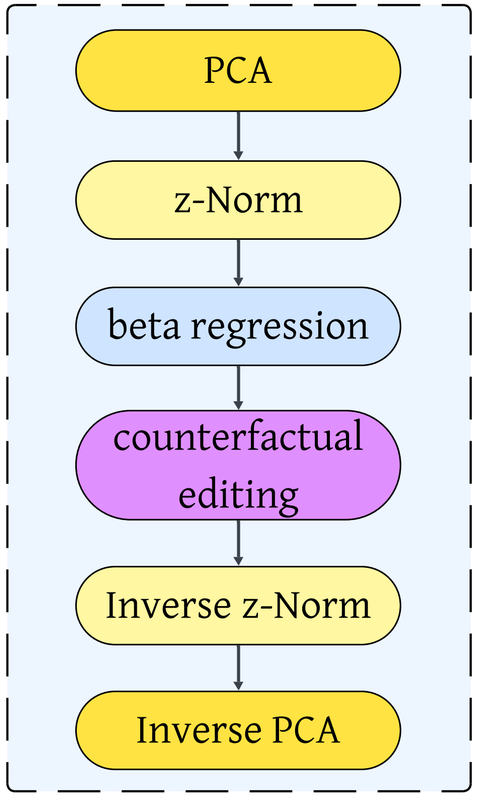


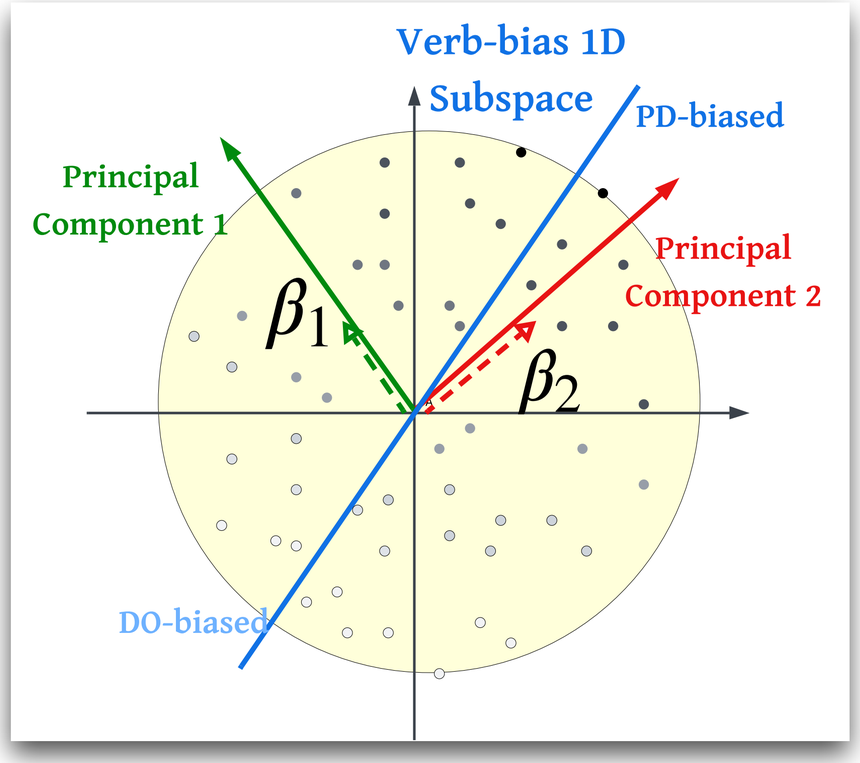


Localizing a 1-dimensional space that encodes verb bias information

[For binary/discrete variable intervention, see Ravfogel et al. 2021, Hao & Linzen 2023, Ozaki et al. 2025]

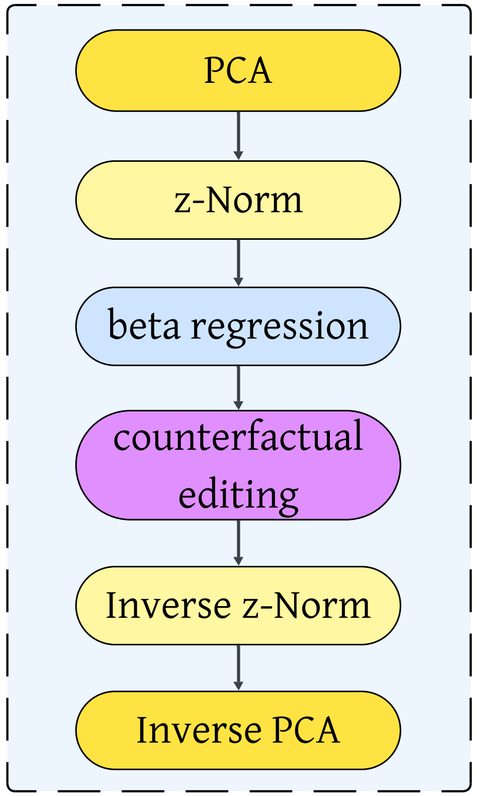
A Method for Continuous Variable Intervention

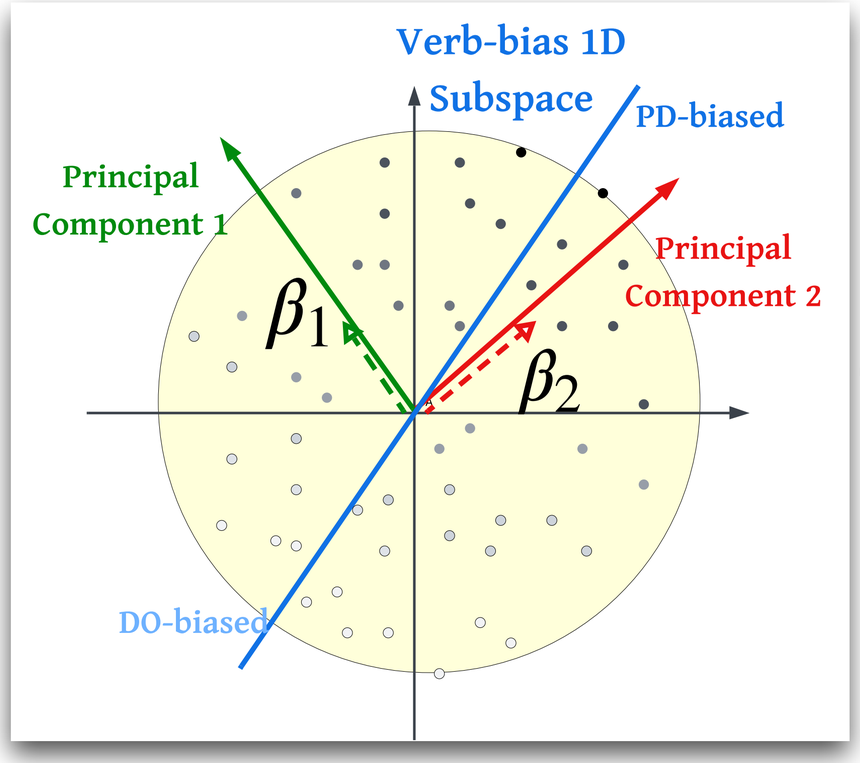


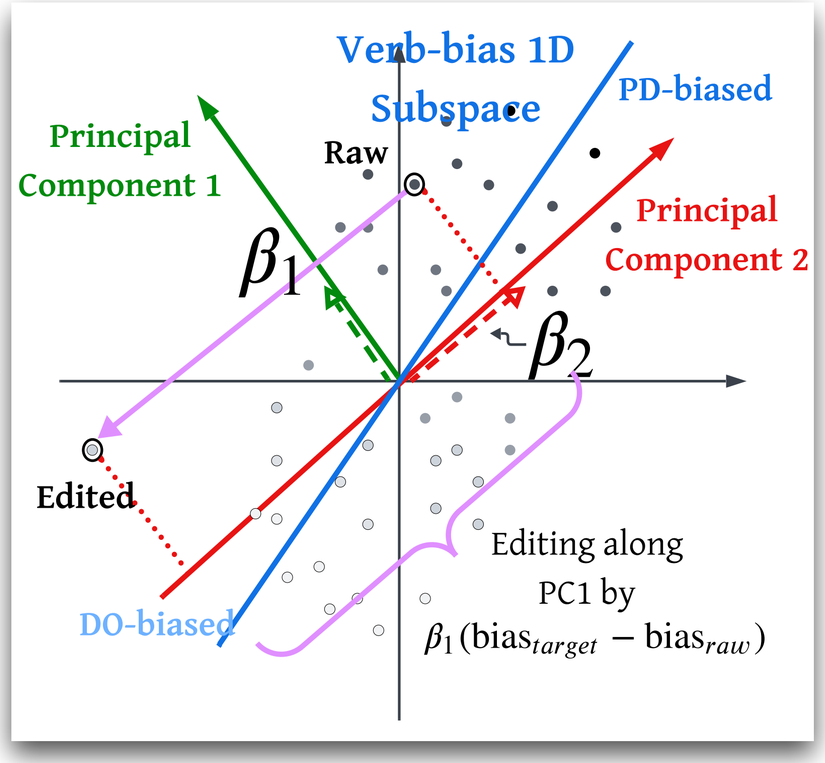


[For a close-formed solution for concept erasing, see Belrose et al. 2023, LEACE]

A Method for Continuous Variable Intervention

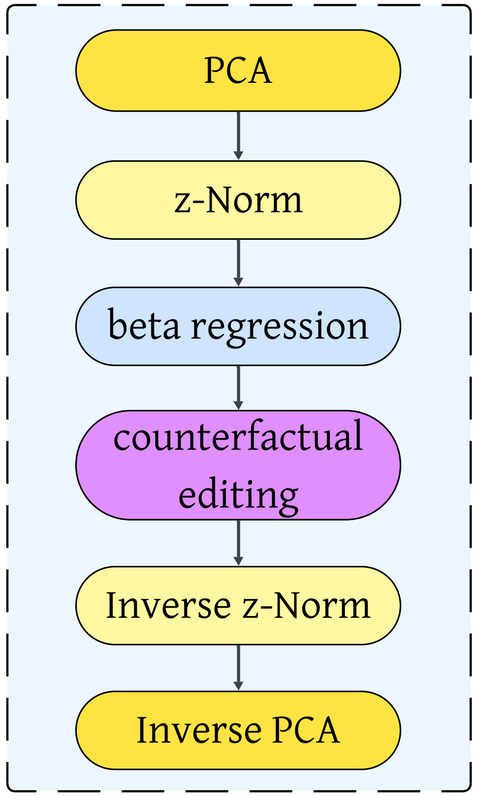


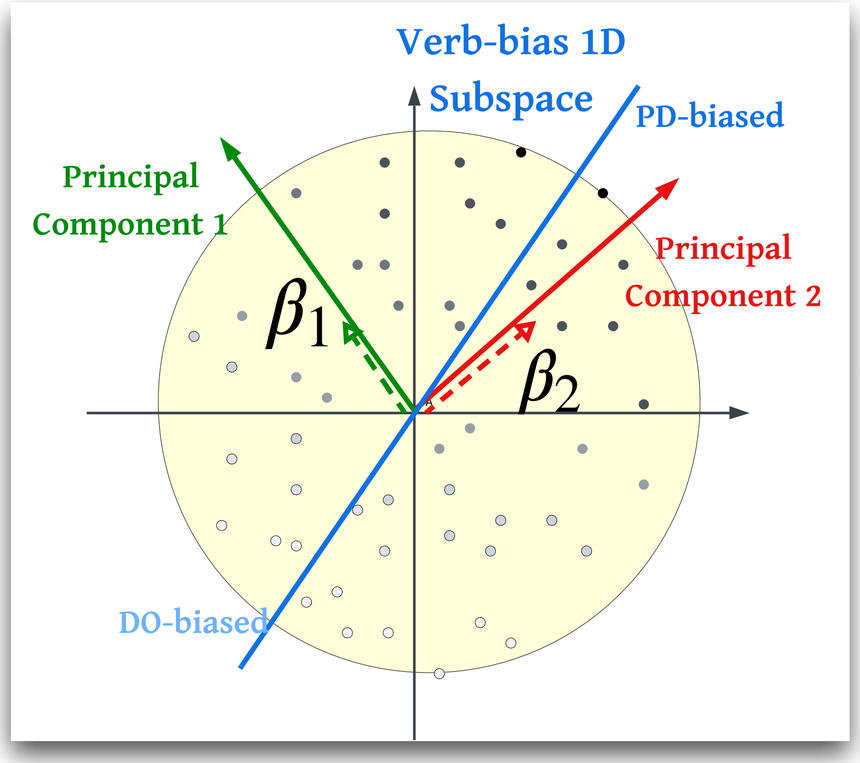


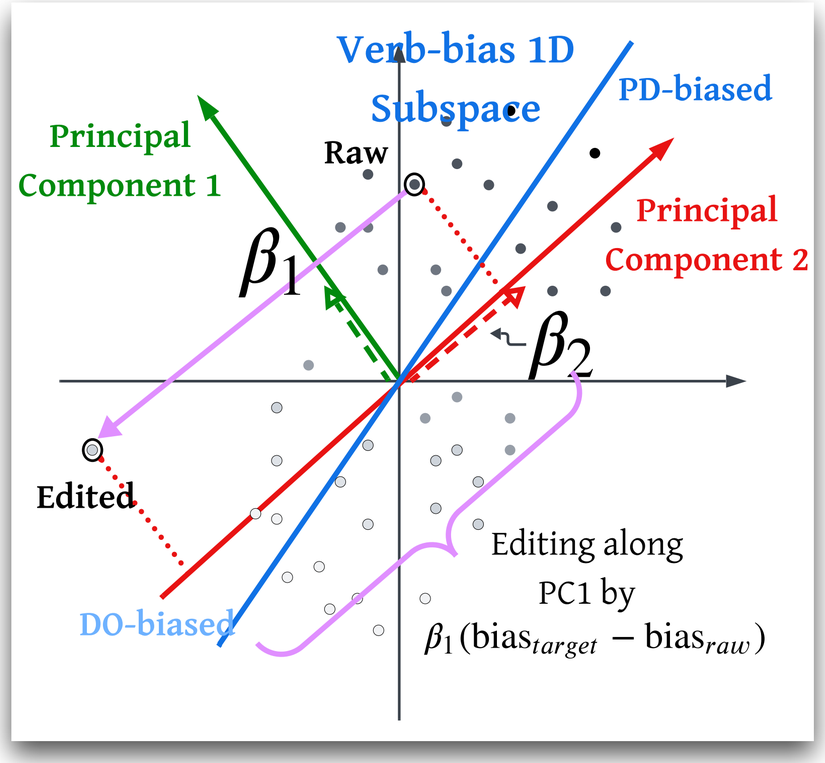


[For a close-formed solution for concept erasing, see Belrose et al. 2023, LEACE]

A Method for Continuous Variable Intervention



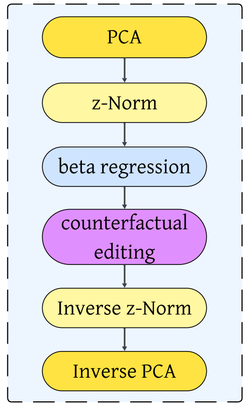




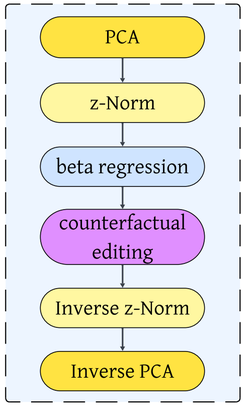
Causally manipulate the verb bias values without changing other information.

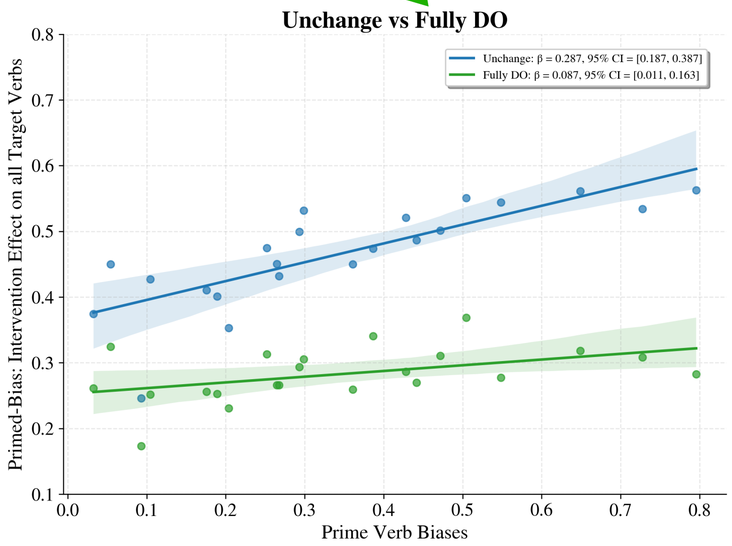
[For a close-formed solution for concept erasing, see Belrose et al. 2023, LEACE]

Causal Verification

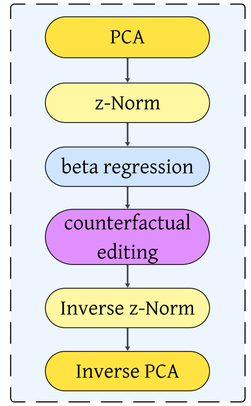


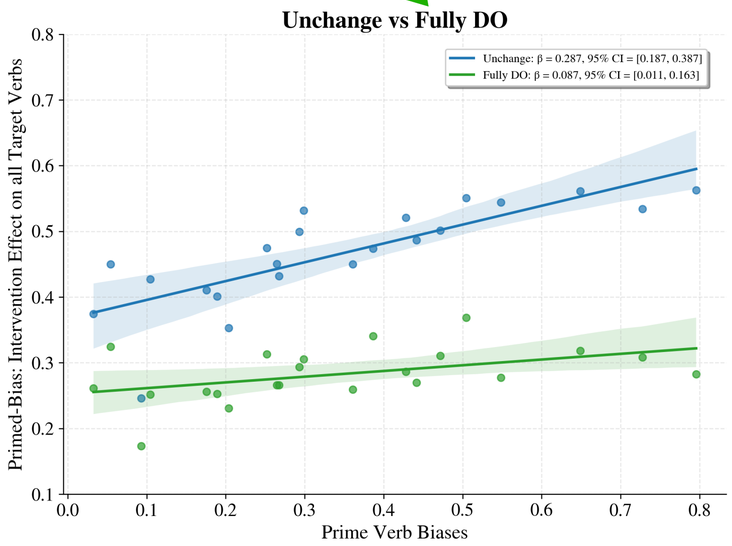
Causal Verification

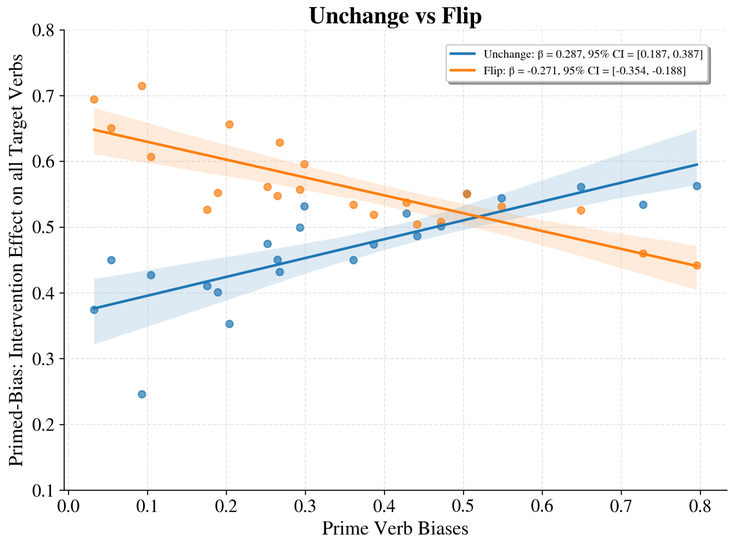




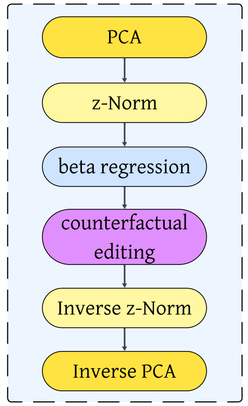
Causal Verification

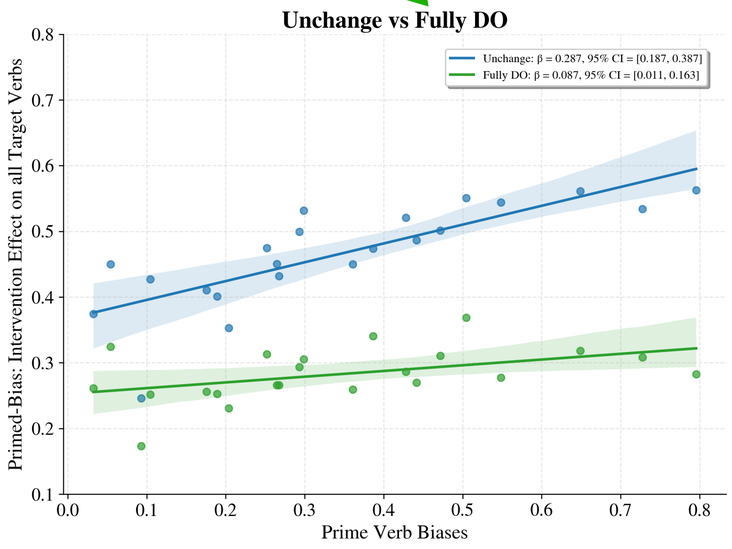


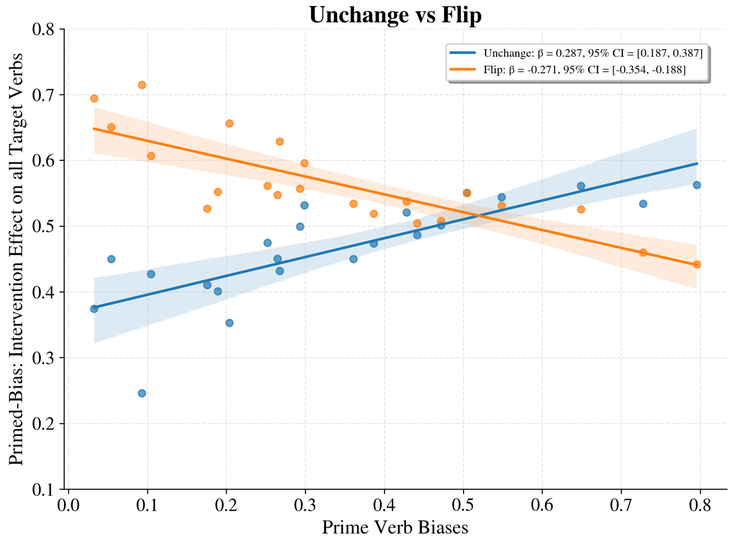


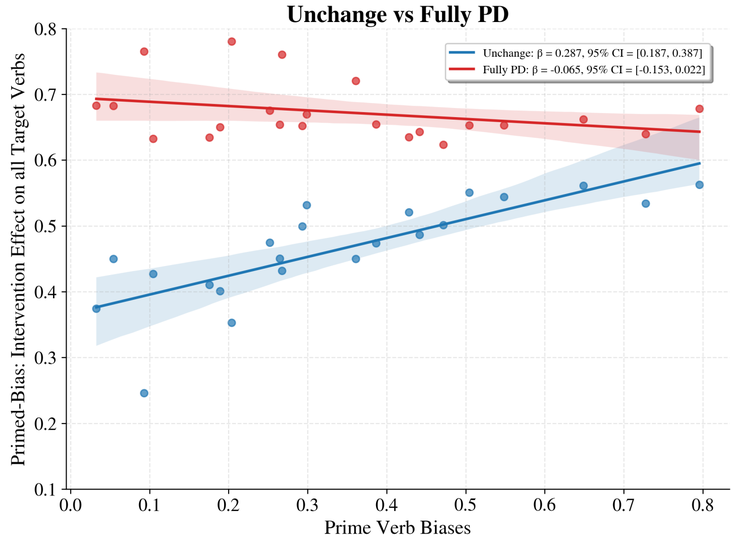


Causal Verification

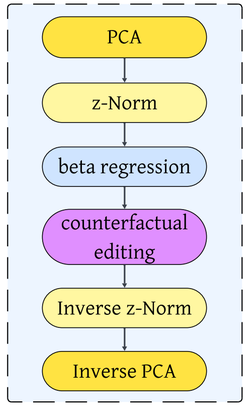


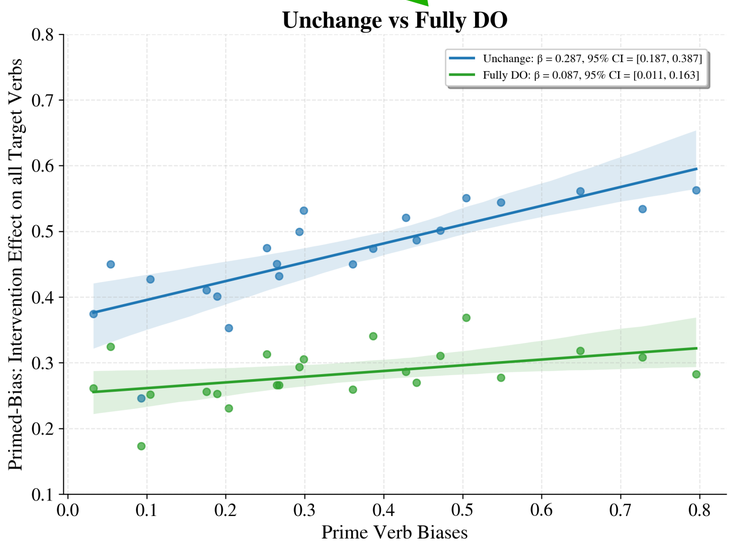


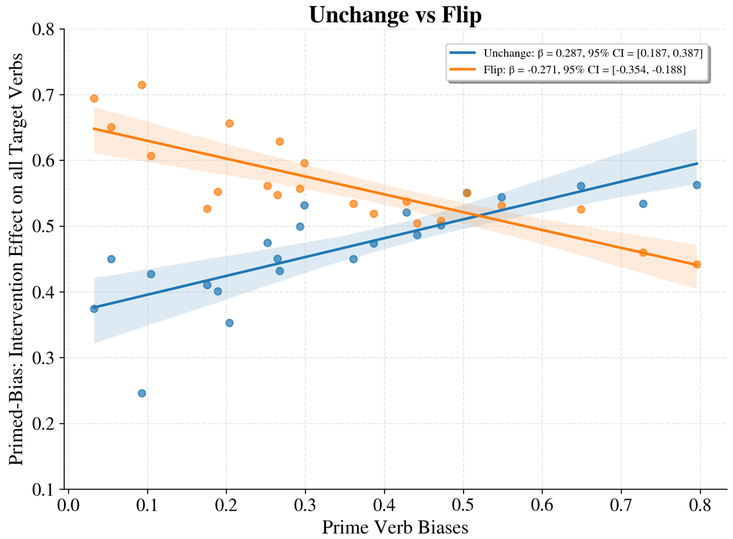


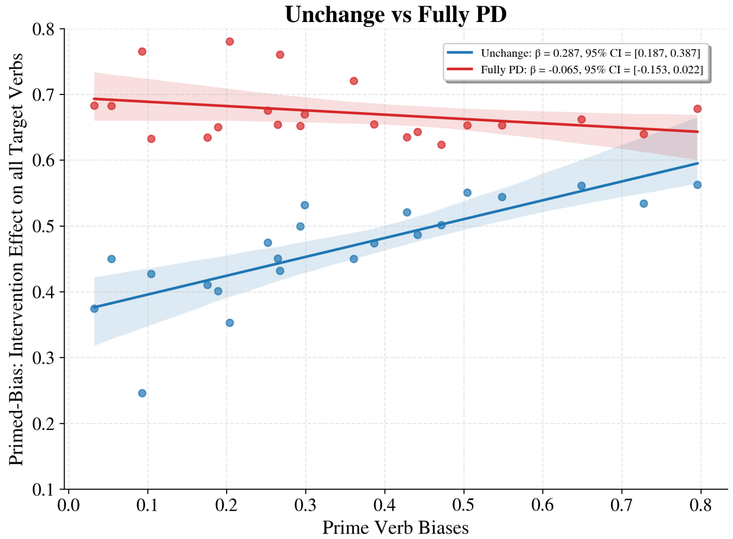


Causal Verification









We showed that manipulating the verb biases causes predicted changes in the priming strength;

⇒ The verb bias representation is causally involved in LLM's structural prediction!

Takeaways from Function Vector Studies

Takeaways from Function Vector Studies

- We use function vectors to further assess the “priming as ICL” claim:
  - **Yes, function vector do lead to standard priming effect**

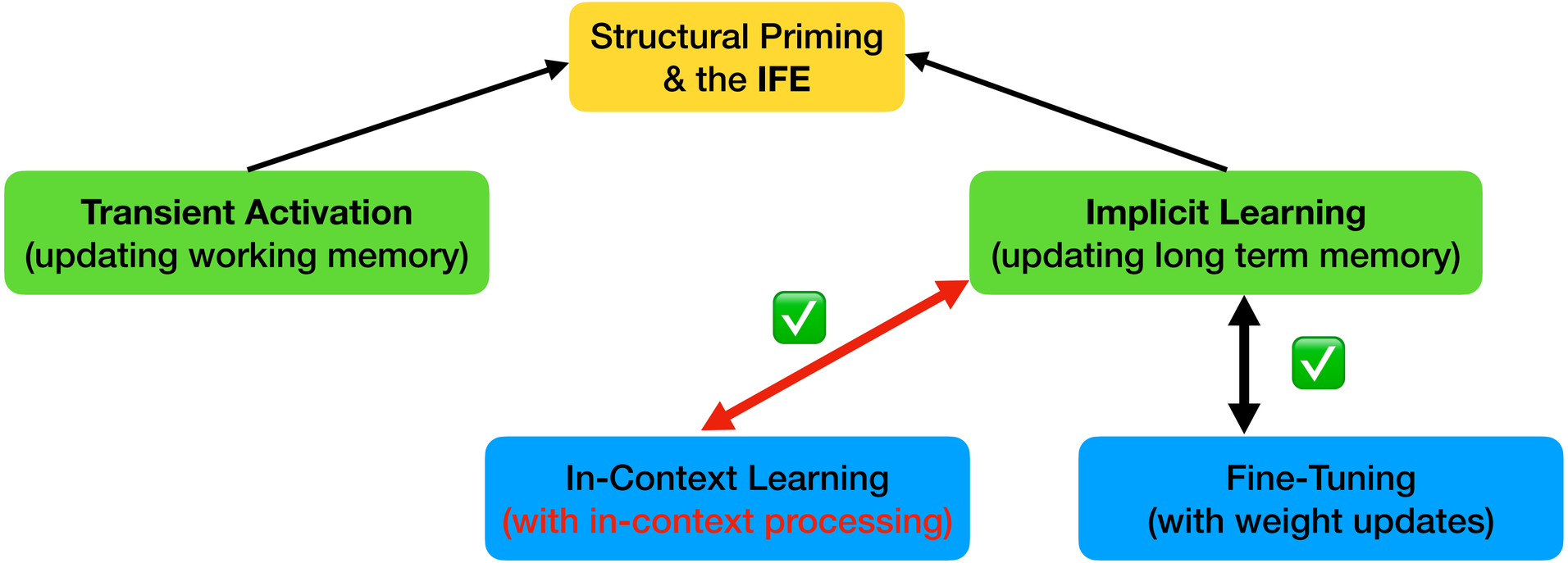
Takeaways from Function Vector Studies

- We use function vectors to further assess the “priming as ICL” claim:
  - **Yes, function vector do lead to standard priming effect**
- What abstract properties are encoded in the function vectors for priming?
  - **We found a low-dimensional subspace that encode verb bias;**
  - **Editing verb bias causally produces predicted effect on verb-specific priming magnitude;**

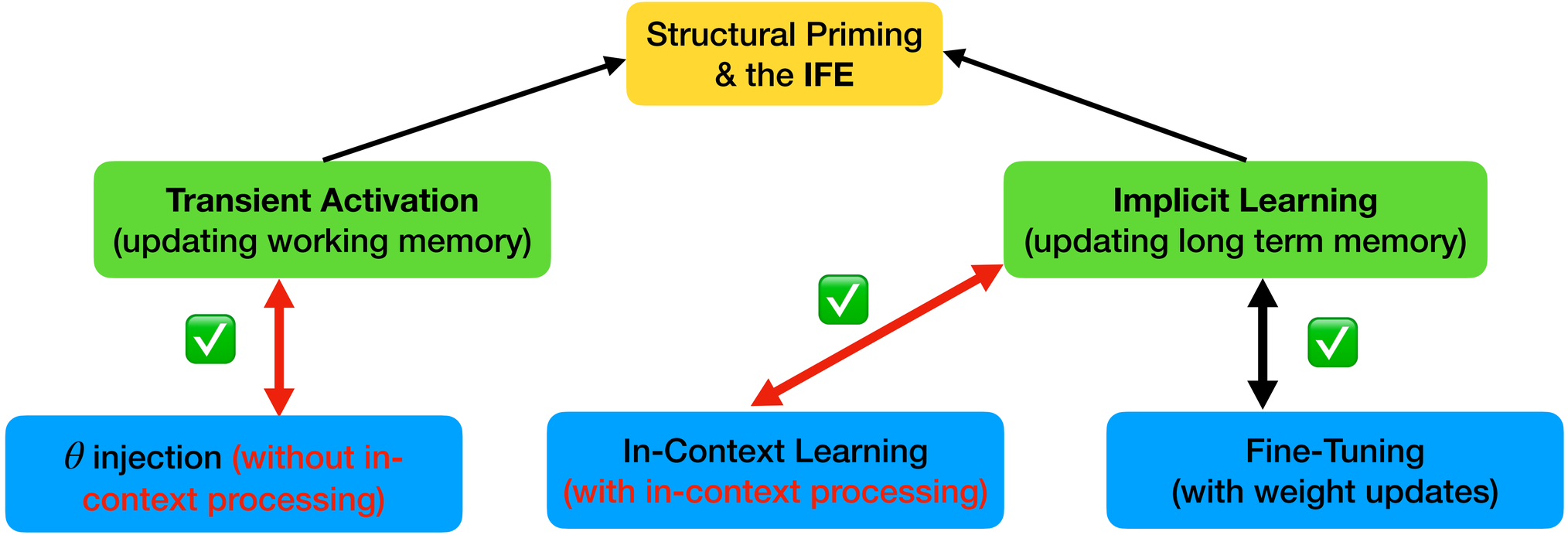
Takeaways from Function Vector Studies

- We use function vectors to further assess the “priming as ICL” claim:
  - **Yes, function vector do lead to standard priming effect**
- What abstract properties are encoded in the function vectors for priming?
  - **We found a low-dimensional subspace that encode verb bias;**
  - **Editing verb bias causally produces predicted effect on verb-specific priming magnitude;**
- However… function vector cannot be the whole story of ICL because:
  - **We see no IFE with the function vector induced priming effect;**
  - **What is the crucial ingredient about IFE that function vectors are missing?**

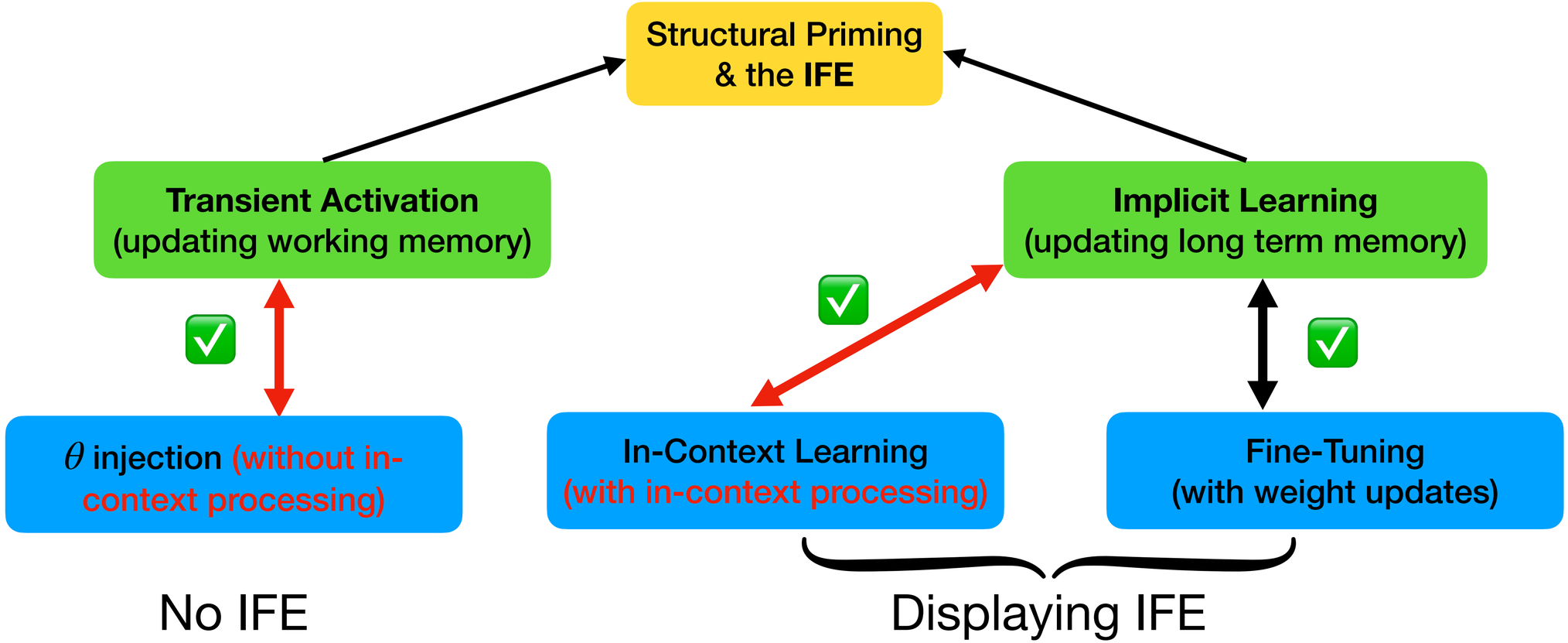
Revisiting the big picture so far



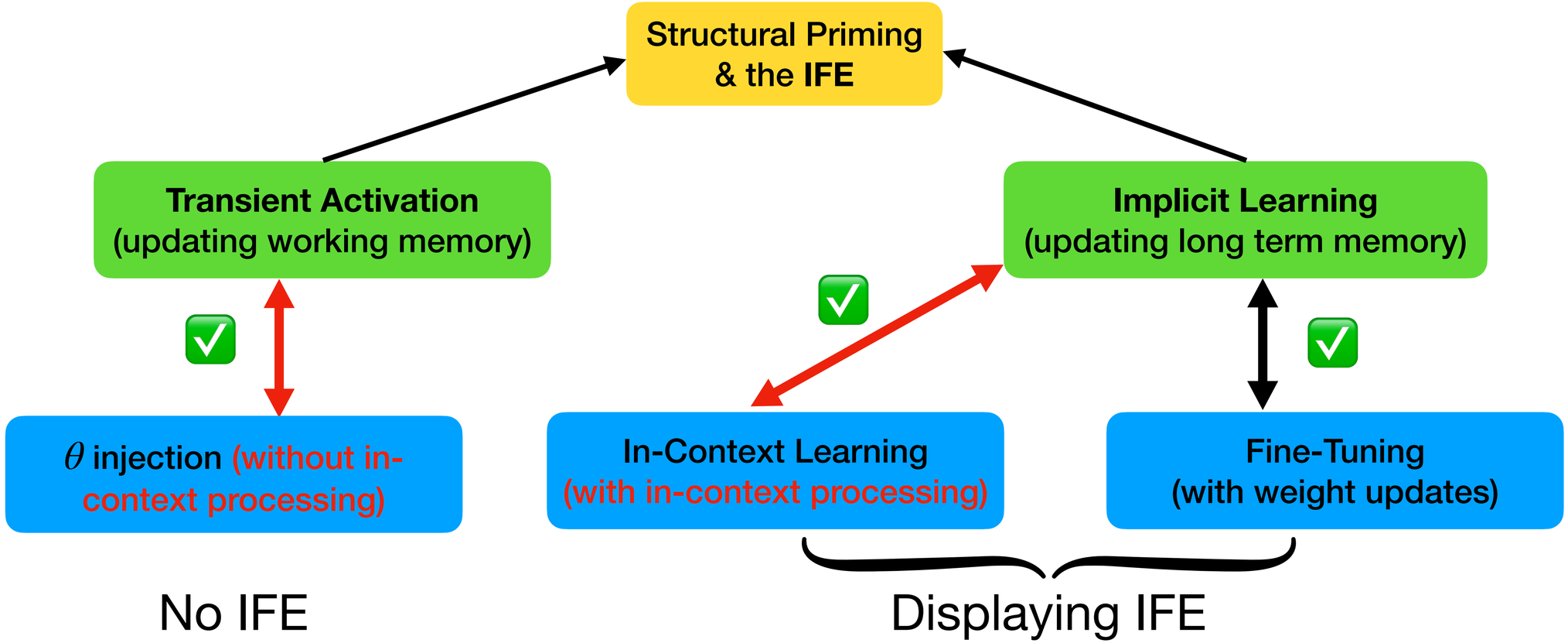
Revisiting the big picture so far



Revisiting the big picture so far



Revisiting the big picture so far



Function vectors do not capture the crucial error signal for LLMs to adapt to context, making it resembling transient activation instead of implicit learning!

High-level Takeaways

What do we learn about ICL from human priming theories?

High-level Takeaways

What do we learn about ICL from human priming theories?

- Previous / popular definition: ICL takes the form of <demo, answer> pairs;

High-level Takeaways

What do we learn about ICL from human priming theories?

- Previous / popular definition: ICL takes the form of <demo, answer> pairs;
- The IFE provides a finer-grained characterization of ICL:
  - Instead of just identifying the “static” task (e.g., antonym, translation), LLMs also dynamically adapt to the context in an error-driven way.

High-level Takeaways

What do we learn about ICL from human priming theories?

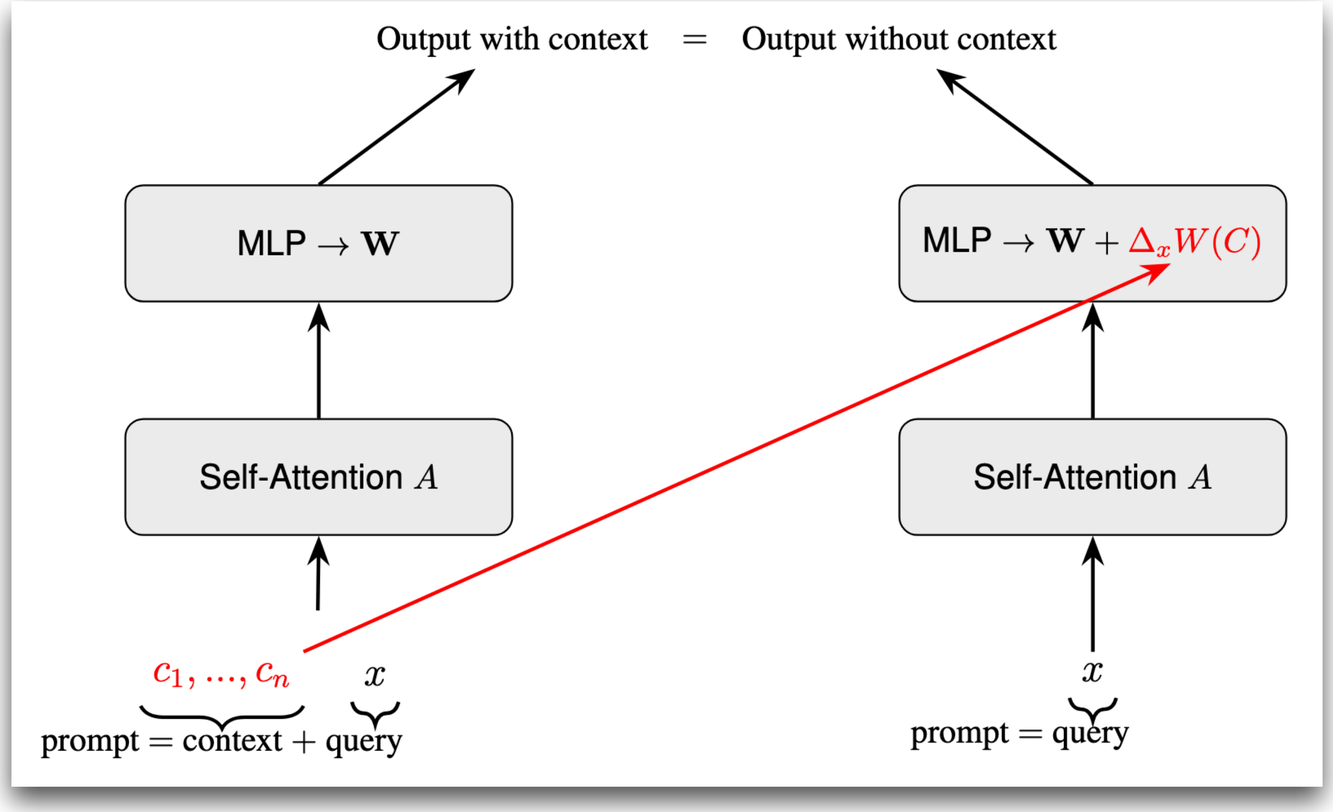
- Previous / popular definition: ICL takes the form of <demo, answer> pairs;
- The IFE provides a finer-grained characterization of ICL:
  - Instead of just identifying the “static” task (e.g., antonym, translation), LLMs also dynamically adapt to the context in an error-driven way.
- This gives rise to a generalized notion of ICL:
  - When we get rid of the “input-output” task form, then ICL becomes the more general ability of “adapting to context” — being context-sensitive in an error-driven way — e.g., Bayesian mind instantiated as predictive coding!

High-level Takeaways

What do we learn about ICL from human priming theories?

- Previous / popular definition: ICL takes the form of <demo, answer> pairs;
- The IFE provides a finer-grained characterization of ICL:
  - Instead of just identifying the “static” task (e.g., antonym, translation), LLMs also dynamically adapt to the context in an error-driven way.
- This gives rise to a generalized notion of ICL:
  - When we get rid of the “input-output” task form, then ICL becomes the more general ability of “adapting to context” — being context-sensitive in an error-driven way — e.g., Bayesian mind instantiated as predictive coding!
- Function vectors only captures the “static” part of ICL — this compression of context is indeed lossy.

Open Questions



Learning without training:
The implicit dynamics of in-context learning

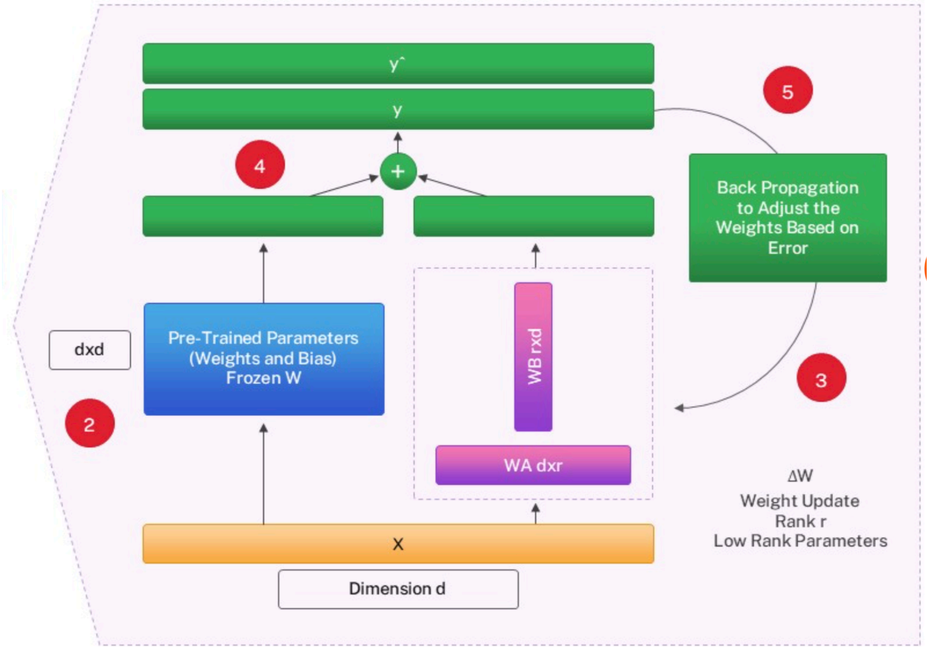
Benoit Dherin* Google Research dherin@google.com

Michael Munn* Google Research munn@google.com

Hanna Mazzawi* Google Research mazzawi@google.com

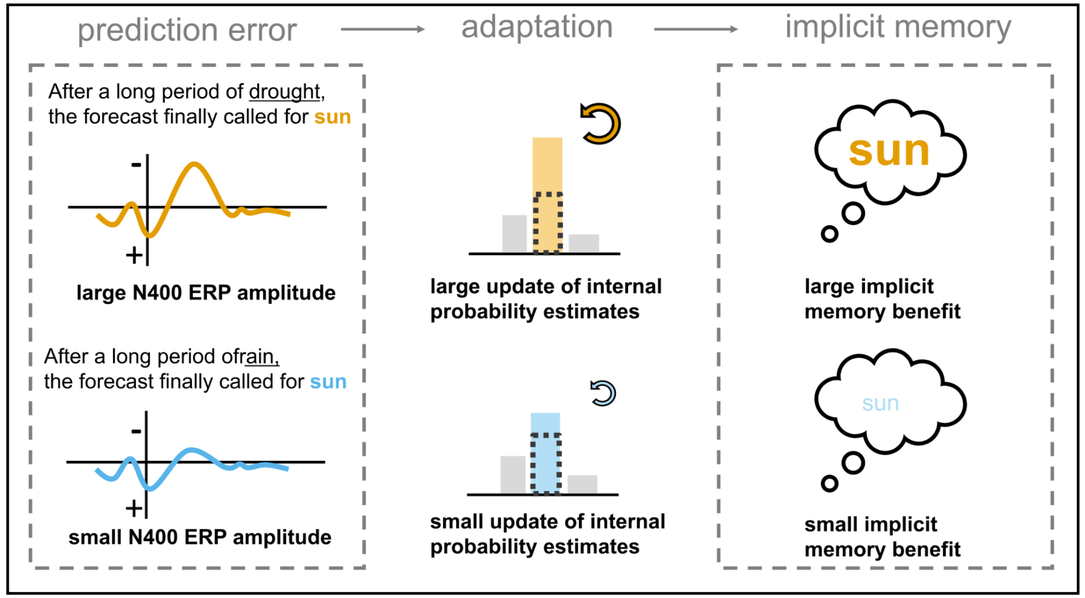
Michael Wunder Google Research mwunder@google.com

Javier Gonzalvo Google Research xavigonzalvo@google.com



Latest progress: contextual influence as implicit low-rank update on MLP weights?

Open Questions: N400 as error signal?



Open Questions: N400 as error signal?

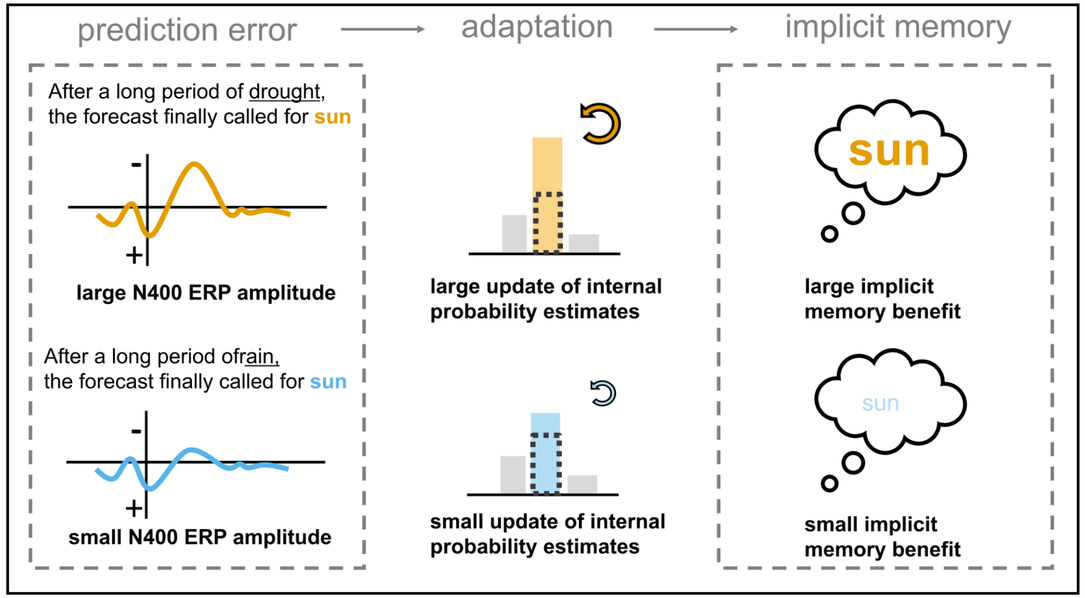
SPECIAL ISSUE:
Cognitive Computational Neuroscience of Language

Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects

James A. Michaelov[1], Megan D. Bardolph[1], Cyma K. Van Petten[2], Benjamin K. Bergen[1], and Seana Coulson[1]

[1]Department of Cognitive Science, University of California, San Diego, La Jolla, CA, USA
[2]Department of Psychology, Binghamton University, State University of New York, Binghamton, NY, USA

Keywords: distributional semantics, ERPs, N400, neural language models, predictive coding



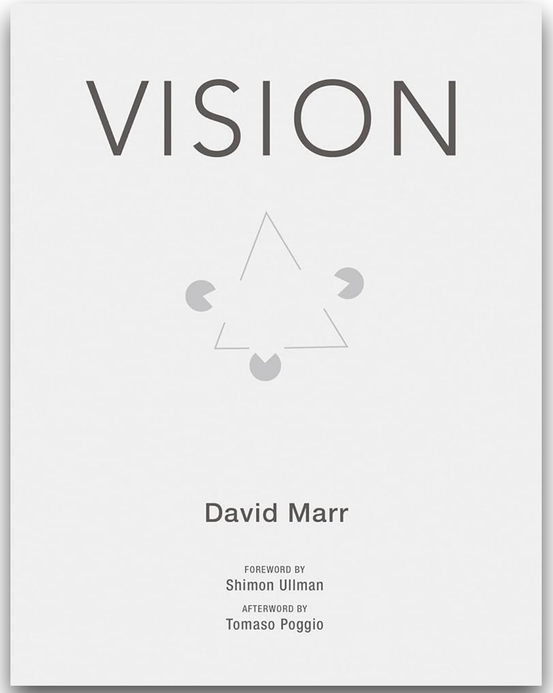
Thanks for Listening!

Q&A Discussion

Appendices

Revisiting Marr's Levels of Analysis

Multiple facets for studying information processing systems

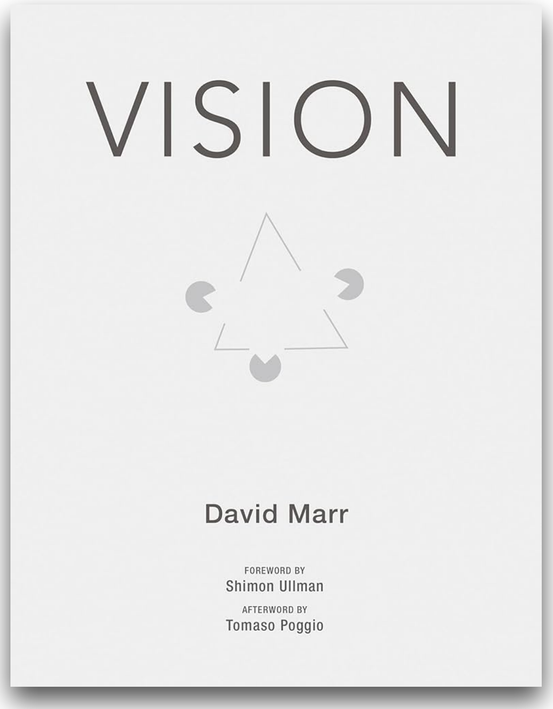


[Marr, 1982]

Revisiting Marr's Levels of Analysis

Multiple facets for studying information processing systems

- **Computational level → [Theoretical linguistics]**
  - [What, why] *Competence, mental grammar, abstract linguistic knowledge*

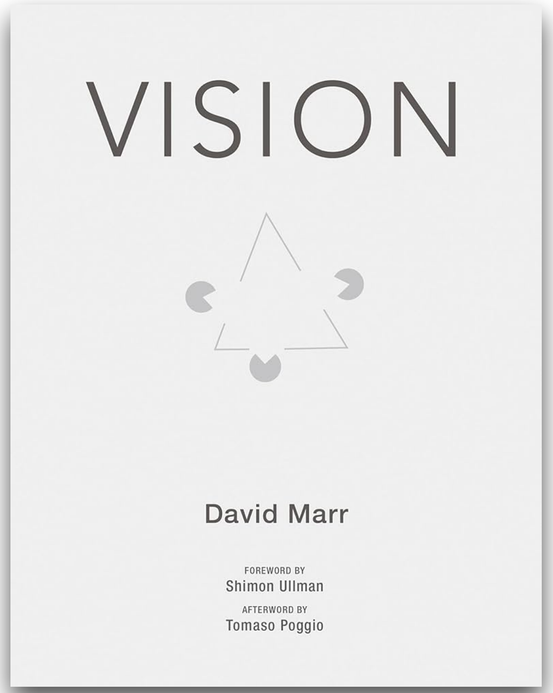


[Marr, 1982]

Revisiting Marr's Levels of Analysis

Multiple facets for studying information processing systems

- **Computational level → [Theoretical linguistics]**
  - [What, why] *Competence, mental grammar, abstract linguistic knowledge*
- **Algorithmic level → [Psycholinguistics]**
  - [How] *Performance, real-time processing*

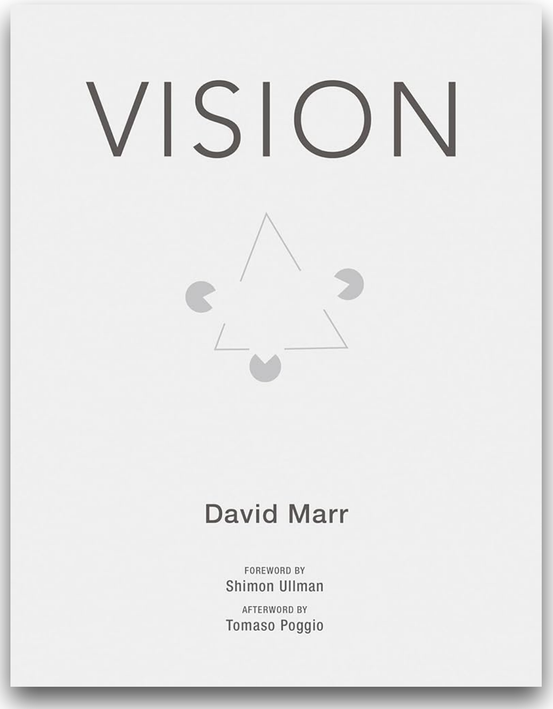


[Marr, 1982]

Revisiting Marr's Levels of Analysis

Multiple facets for studying information processing systems

- **Computational level → [Theoretical linguistics]**
  - [What, why] *Competence, mental grammar, abstract linguistic knowledge*
- **Algorithmic level → [Psycholinguistics]**
  - [How] *Performance, real-time processing*
- **Implementational level → [Neurolinguistics]**
  - [How - lower level] *Performance, neural instantiations of language*



[Marr, 1982]

Current Proposal: The Full Picture

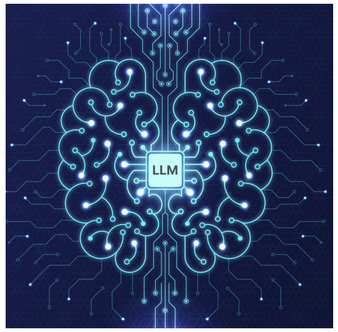
Current Proposal: The Full Picture

Computational Level

Current Proposal: The Full Picture

Computational Level

***Language* in the broader design of intelligent systems**

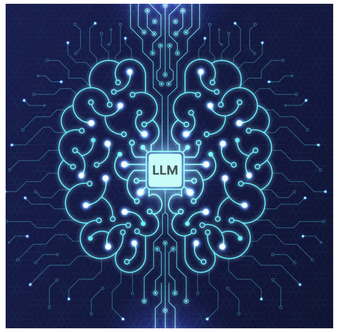


Current Proposal: The Full Picture

Computational Level

***Language* in the broader design of intelligent systems**

Algorithmic Level ONE



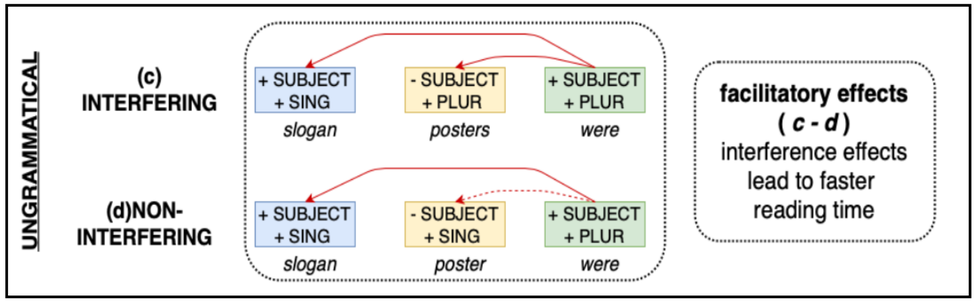
Current Proposal: The Full Picture

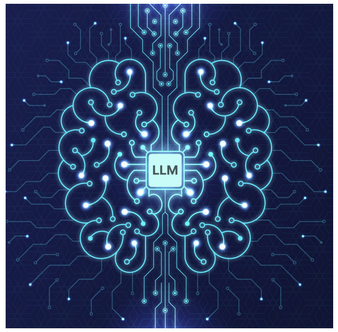
Computational Level

Language **in the broader design of intelligent systems**

Algorithmic Level ONE

Psycholinguistic theories (?)



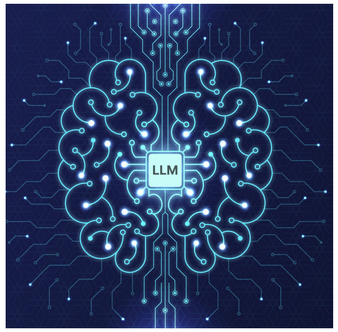


An “algorithm” of sentence processing

Current Proposal: The Full Picture

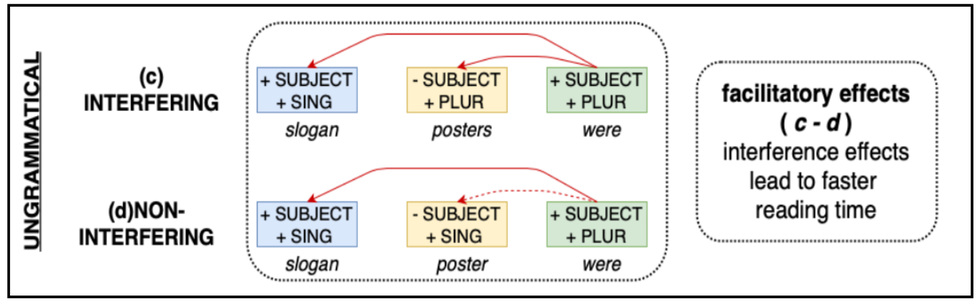
Computational Level

Language **in the broader design of intelligent systems**



Algorithmic Level ONE

Psycholinguistic theories (?)



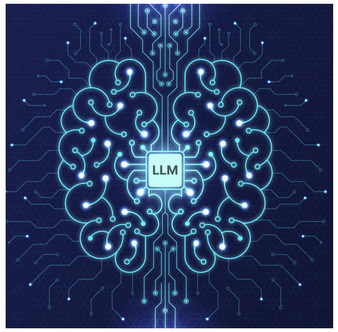
Algorithmic Level TWO

An “algorithm” of sentence processing

Current Proposal: The Full Picture

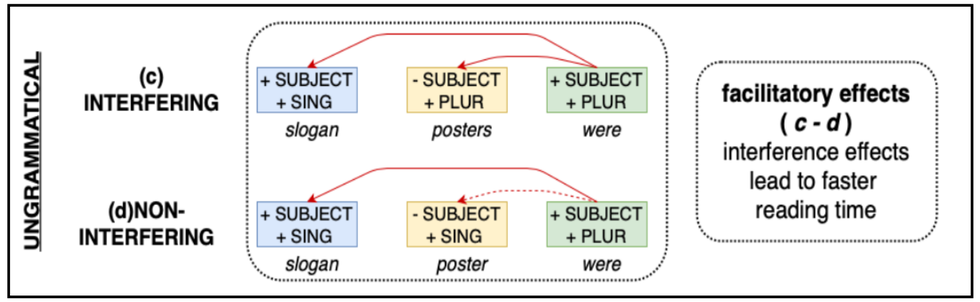
Computational Level

Language **in the broader design of intelligent systems**



Algorithmic Level ONE

Psycholinguistic theories (?)

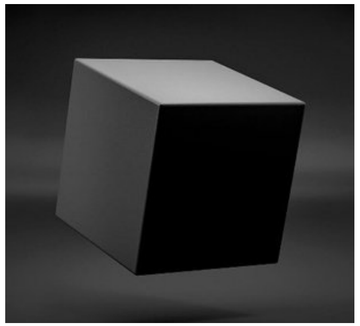


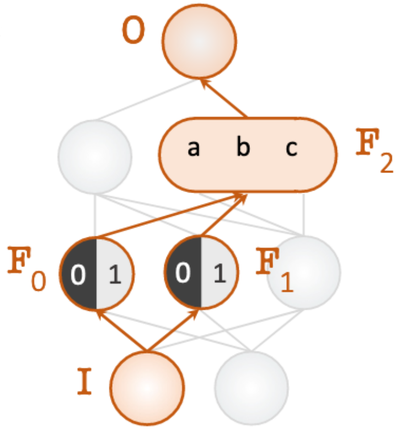
Functionally approximating

Functionally approximating

Algorithmic Level TWO

An "algorithm" of sentence processing





LMs are shaped by what they are trained on!

Data Distributional Properties Drive Emergent In-Context Learning in Transformers

Stephanie C.Y. Chan, DeepMind
Adam Santoro, DeepMind
Andrew K. Lampinen, DeepMind
Jane X. Wang, DeepMind
Aaditya K. Singh, University College London
Pierre H. Richemond, DeepMind
James L. McClelland, DeepMind, Stanford University
Felix Hill, DeepMind

Burstiness as a distributional property!

Compositional Structures!

Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

Sewon Min[1,2] Xinxi Lyu[1] Ari Holtzman[1] Mikel Artetxe[2]
Mike Lewis[2] Hannaneh Hajishirzi[1,3] Luke Zettlemoyer[1,2]

Parallelism!

Parallel Structures in Pre-training Data Yield In-Context Learning

Yanda Chen[1] Chen Zhao[2,3] Zhou Yu[1] Kathleen McKeown[1] He He[2]

⇒ The IFE as a diagnostic of the error-driven learning mechanism in ICL!