Adaptation-to-Context in Humans and Large Language Models

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

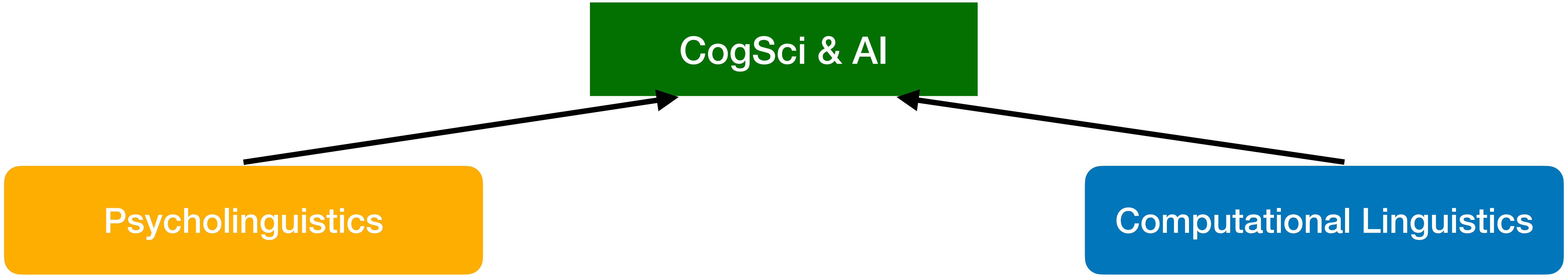
Zhenghao Herbert Zhou 周正浩, Yale University

LA Reading Group, Peking University

December 19, 2025

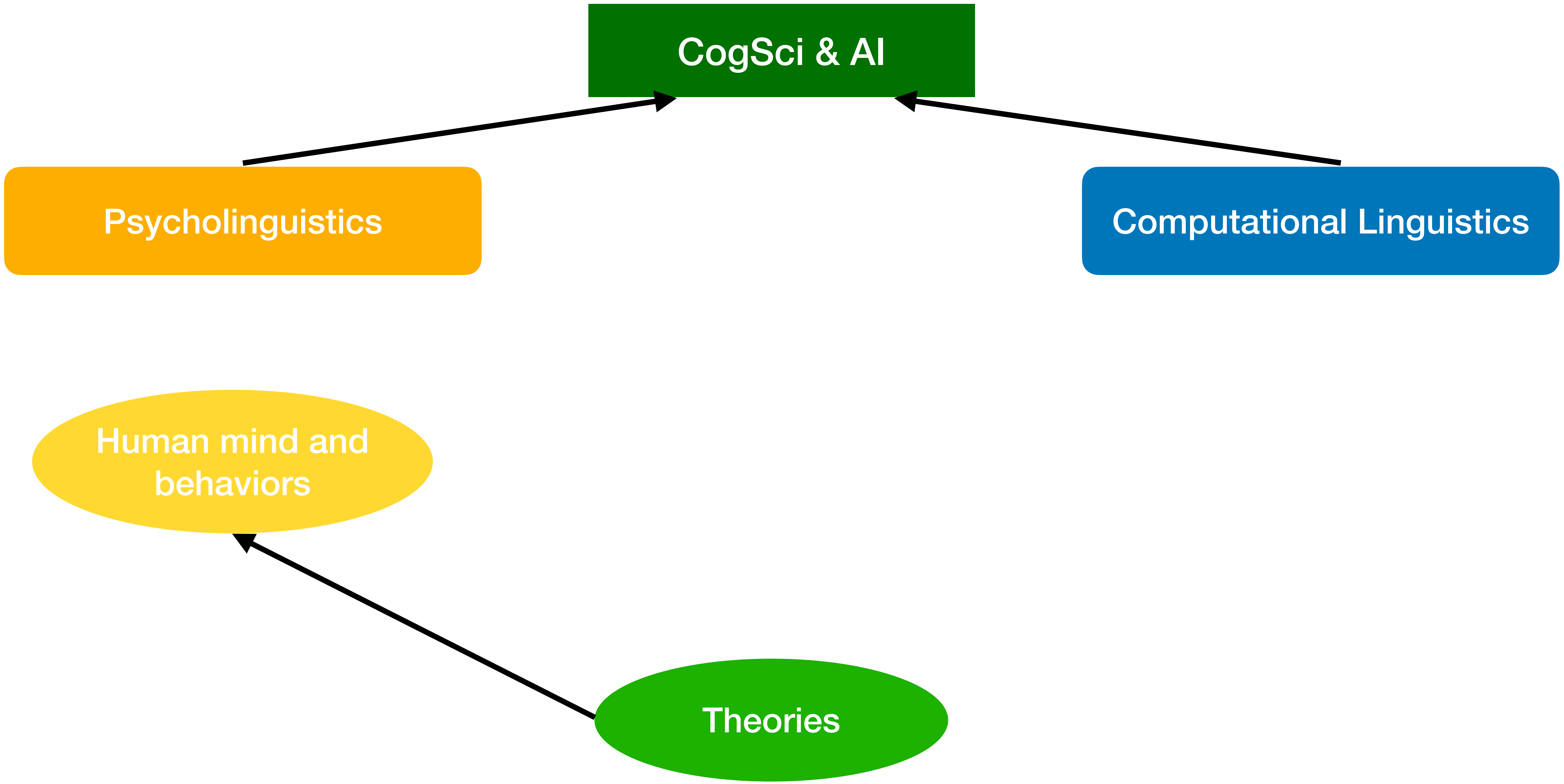
The Big Picture: Notes on Methodology

Humans and LLMs are bidirectionally informative



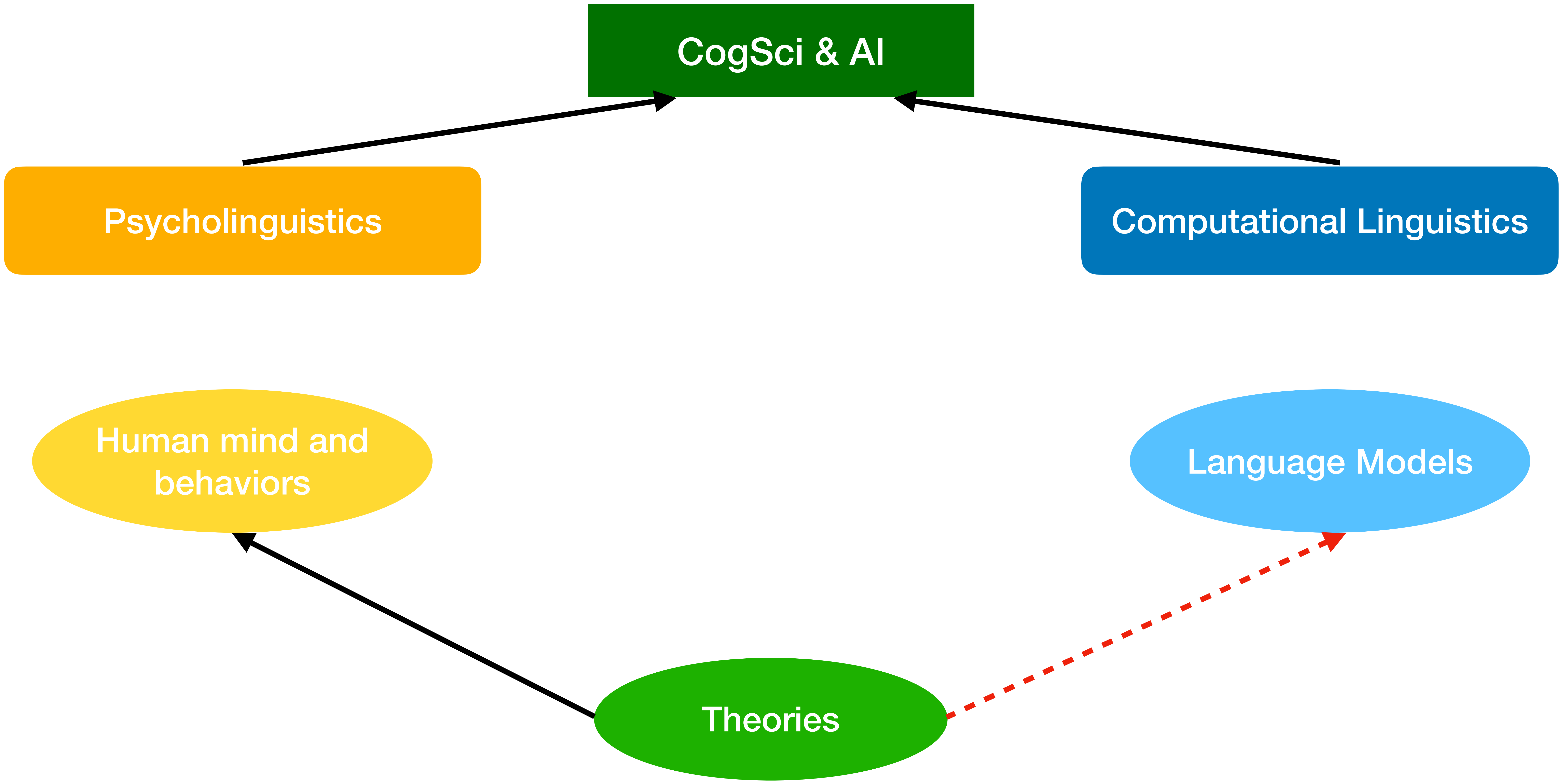
The Big Picture: Notes on Methodology

Humans and LLMs are bidirectionally informative



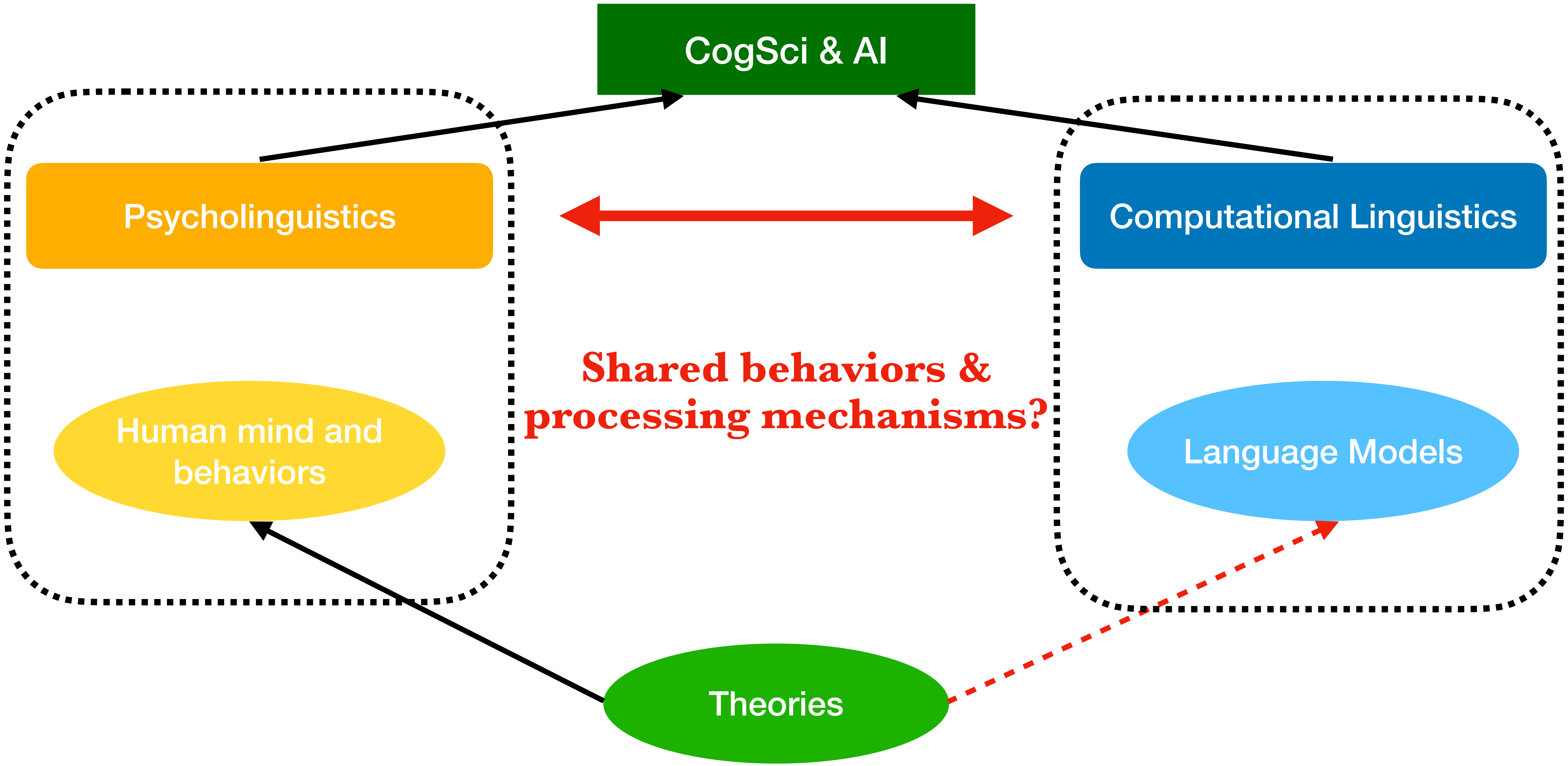
The Big Picture: Notes on Methodology

Humans and LLMs are bidirectionally informative



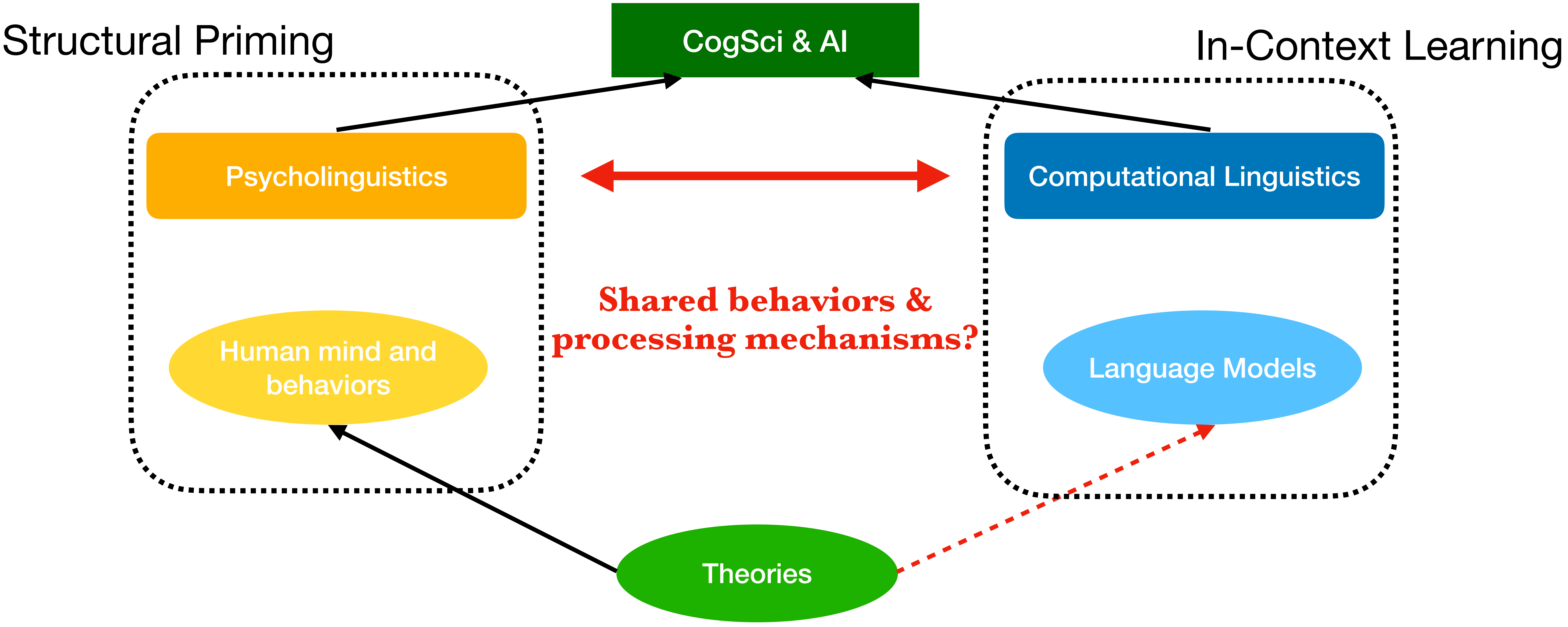
The Big Picture: Notes on Methodology

Humans and LLMs are bidirectionally informative



The Big Picture: Notes on Methodology

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.

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

Verb Bias:

buy is biased towards DO

design is towards PD

A doctor bought a chief a plate. <u>The secretary drew the band the card.</u>

A doctor designed a chief a plate. ***<u>The secretary drew the band the card.</u>***

[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.

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

Verb Bias:

buy is biased towards DO

design is towards PD

A doctor bought a chief a plate. <u>The secretary drew the band the card.</u>

A doctor designed a chief a plate. ***<u>The secretary drew the band the card.</u>***

Greater priming effect!

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

The Inverse Frequency Effect (IFE)

Prime$_{DO}$ → Target$_{DO}$

Unprimed probability = 0.2

A doctor bought a chief a plate. <u>The secretary drew the band the card.</u>

A doctor designed a chief a plate. <u>The secretary drew the band the card.</u>

[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. *<u>The secretary drew the band the card.</u>*

A doctor designed a chief a plate. *<u>The secretary drew the band the card.</u>*

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.

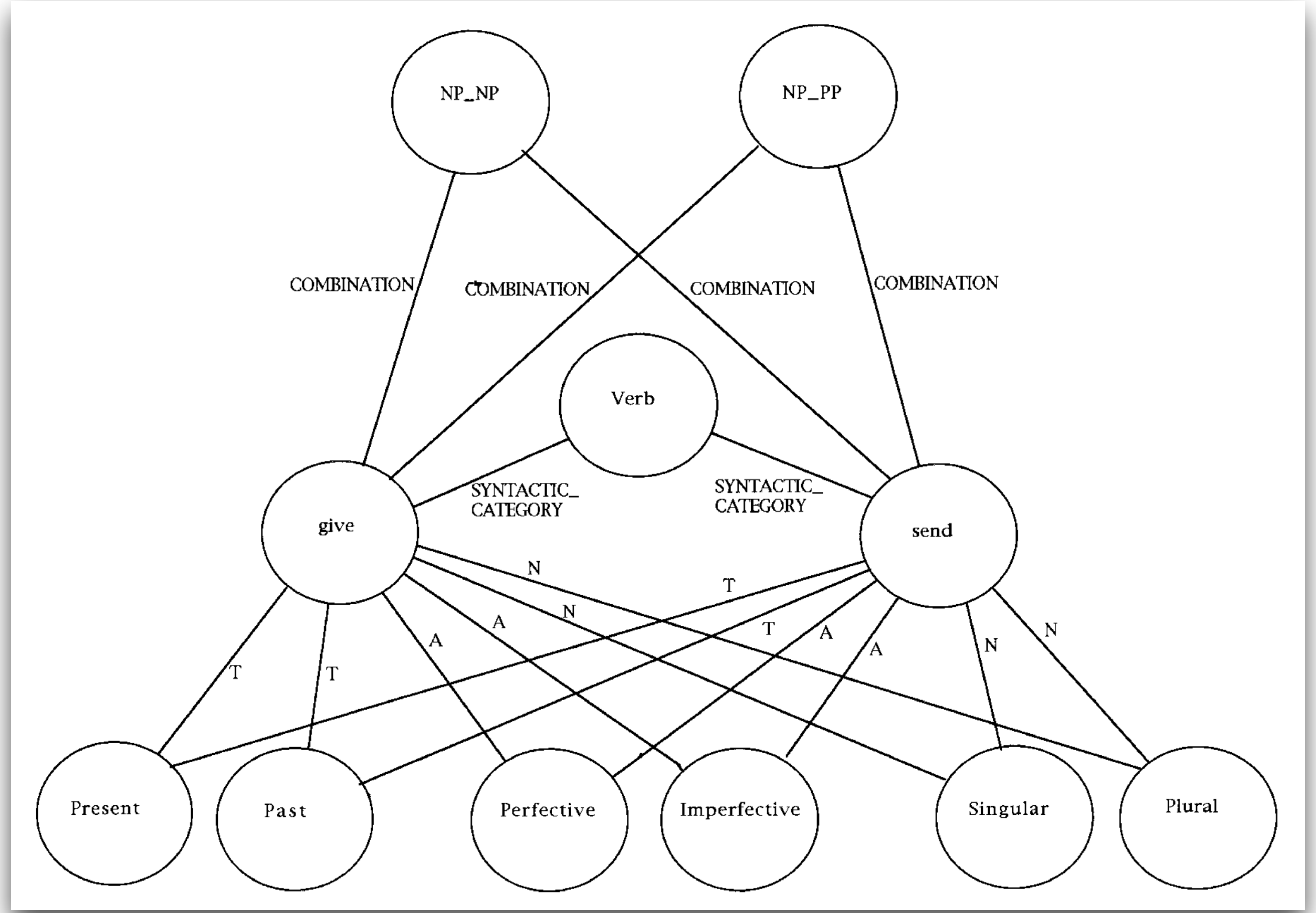
A doctor designed a chief a plate. The secretary drew the band the card.

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

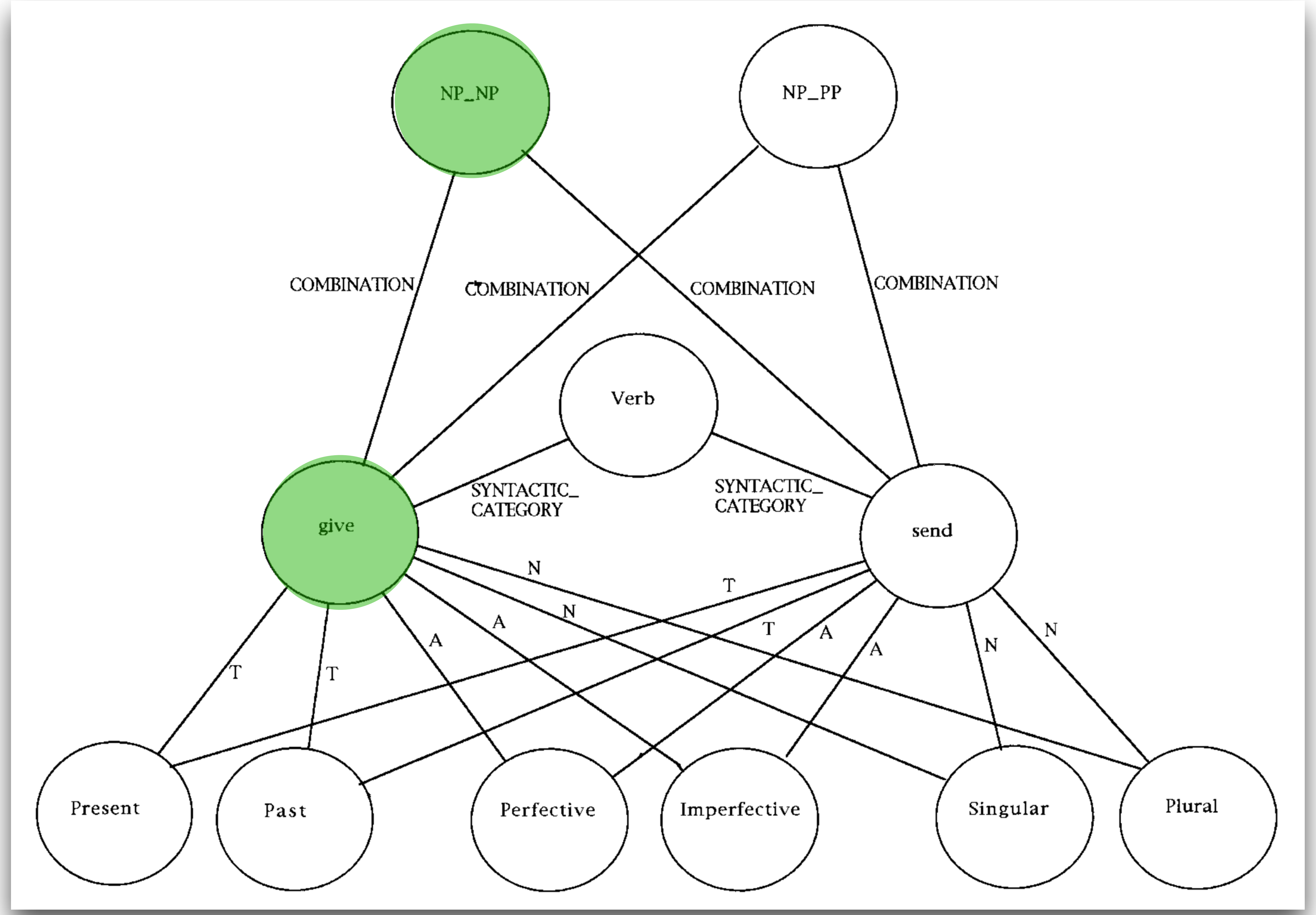


Structural Nodes

Lexical Nodes

[e.g., Pickering & Branigan 1998]

Theory 1: Transient Activation



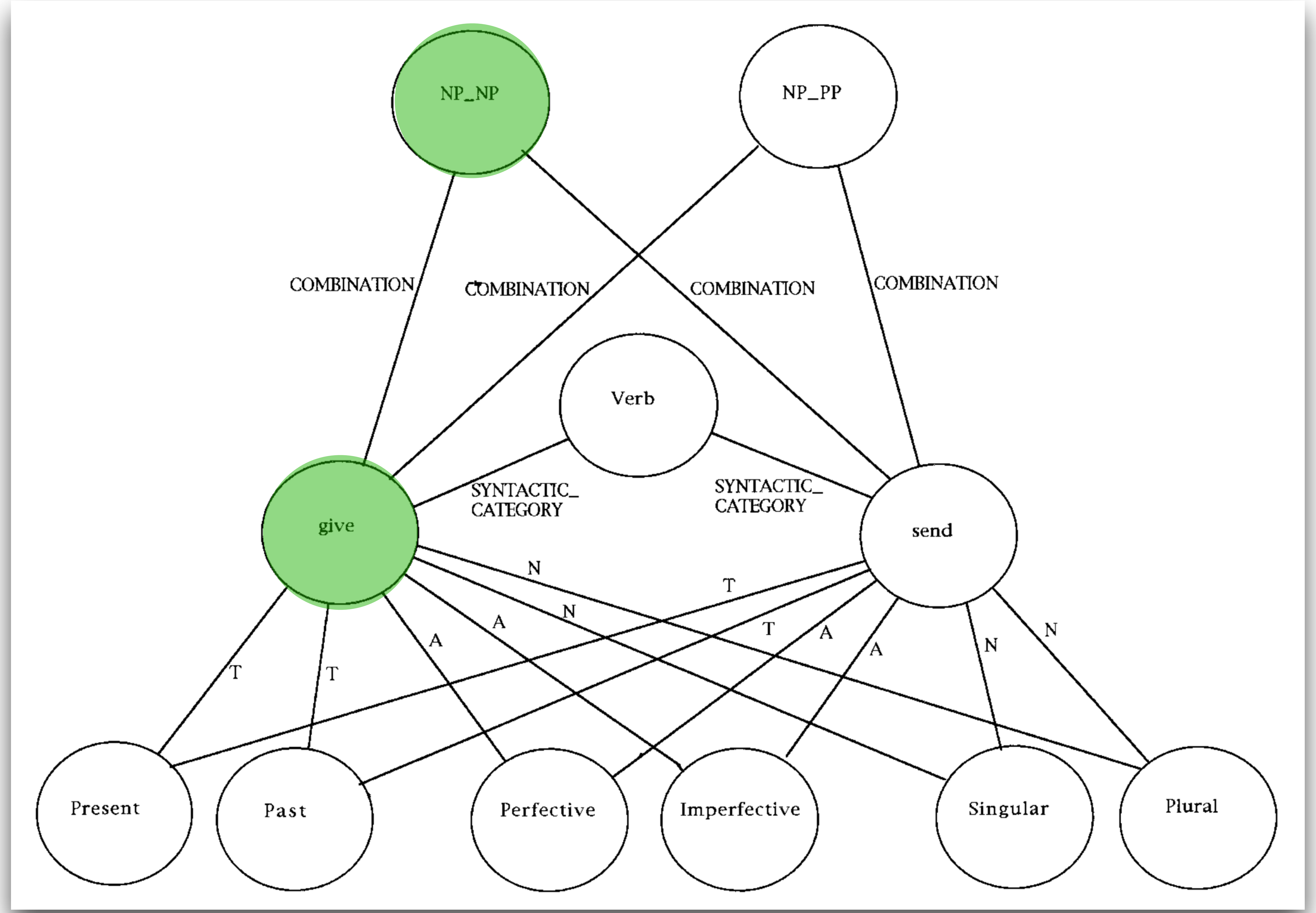
Structural Nodes

Lexical Nodes

[e.g., Pickering & Branigan 1998]

Theory 1: Transient Activation

Quickly decaying activations



Structural Nodes

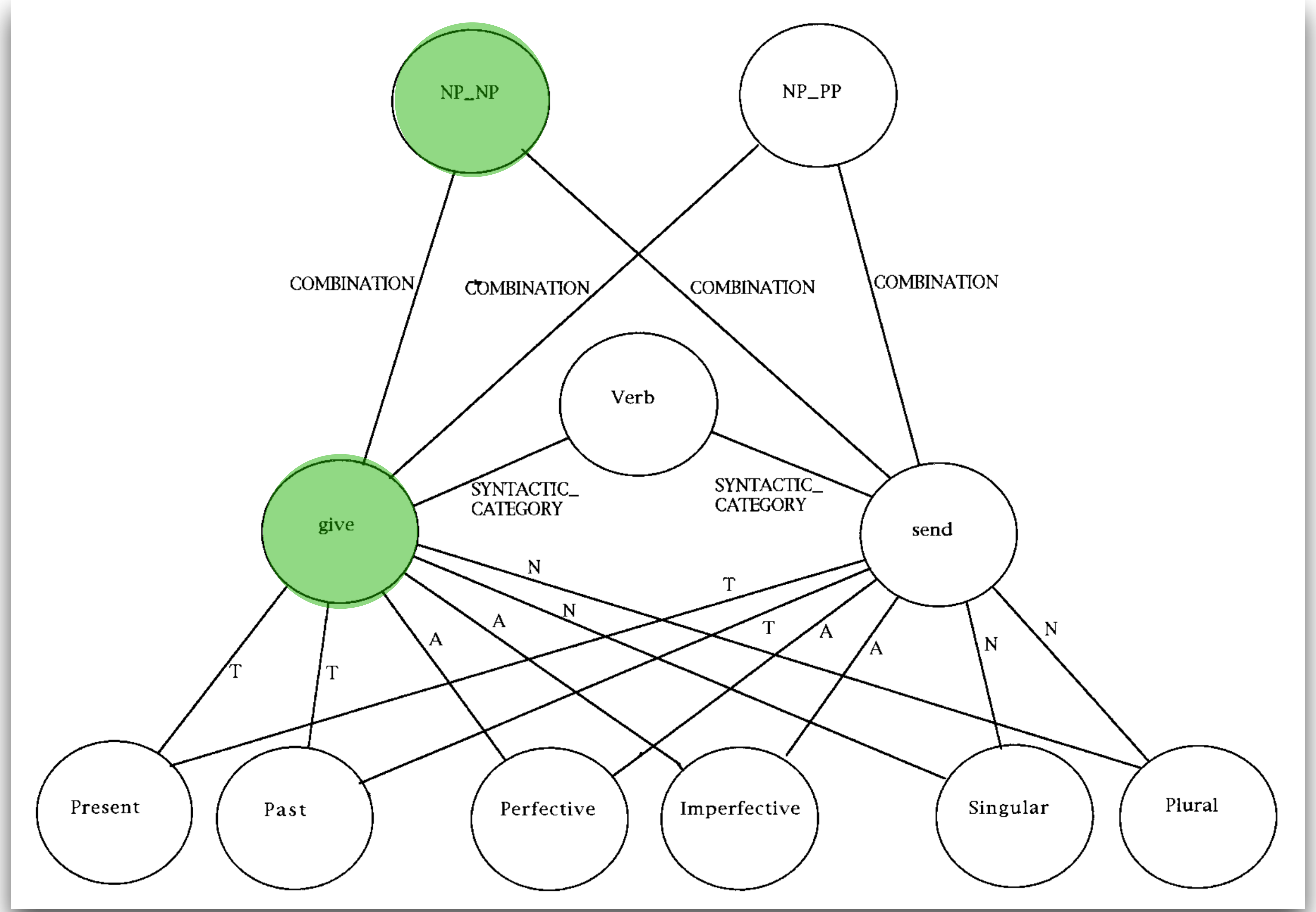
Increasing the activation values of the nodes, which persists through next productions.

Lexical Nodes

[e.g., Pickering & Branigan 1998]

Theory 1: Transient Activation

Quickly decaying activations



Structural Nodes

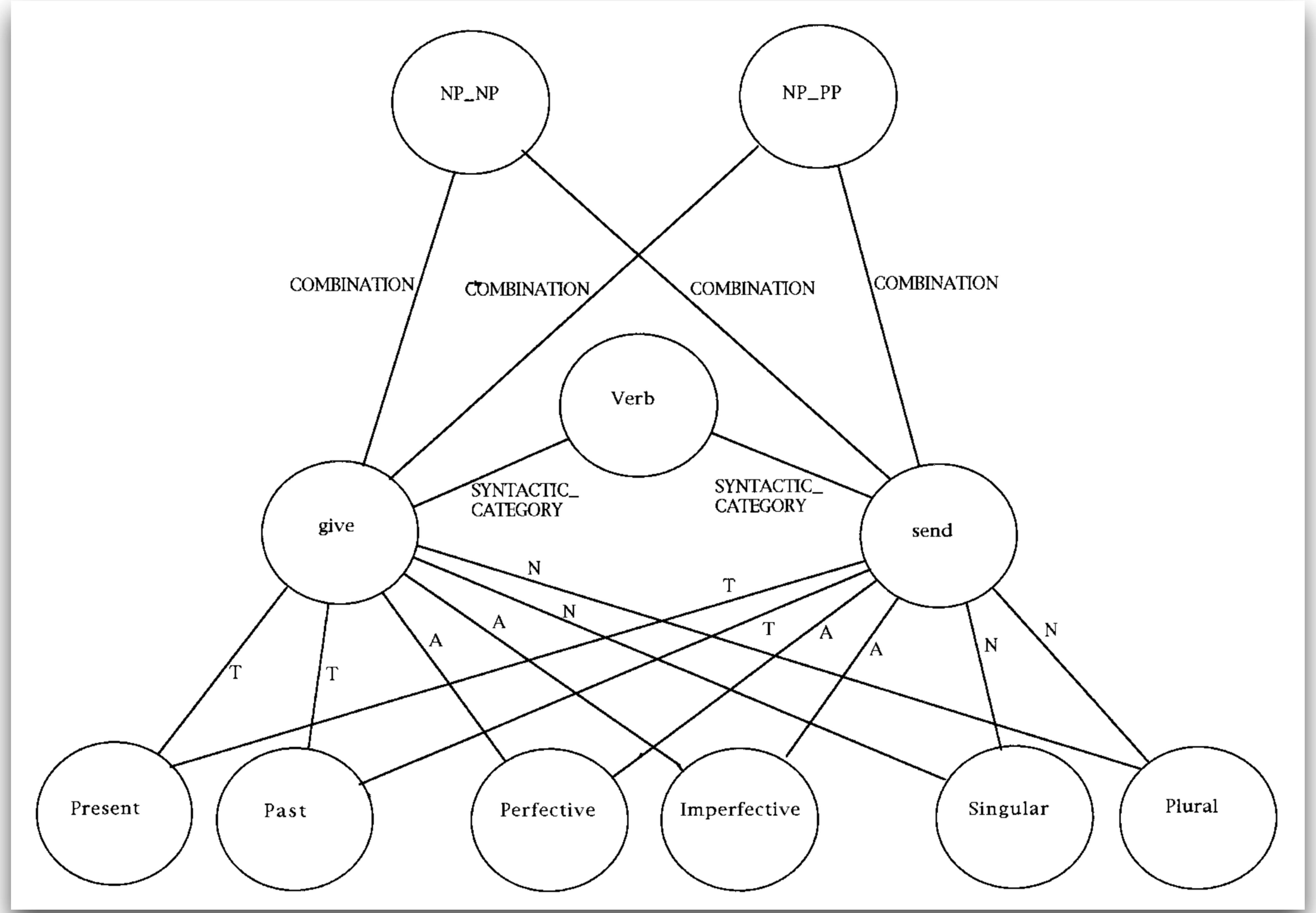
Increasing the activation values of the nodes, which persists through next productions.

Lexical Nodes

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

[e.g., Pickering & Branigan 1998]

Theory 2: Implicit Learning

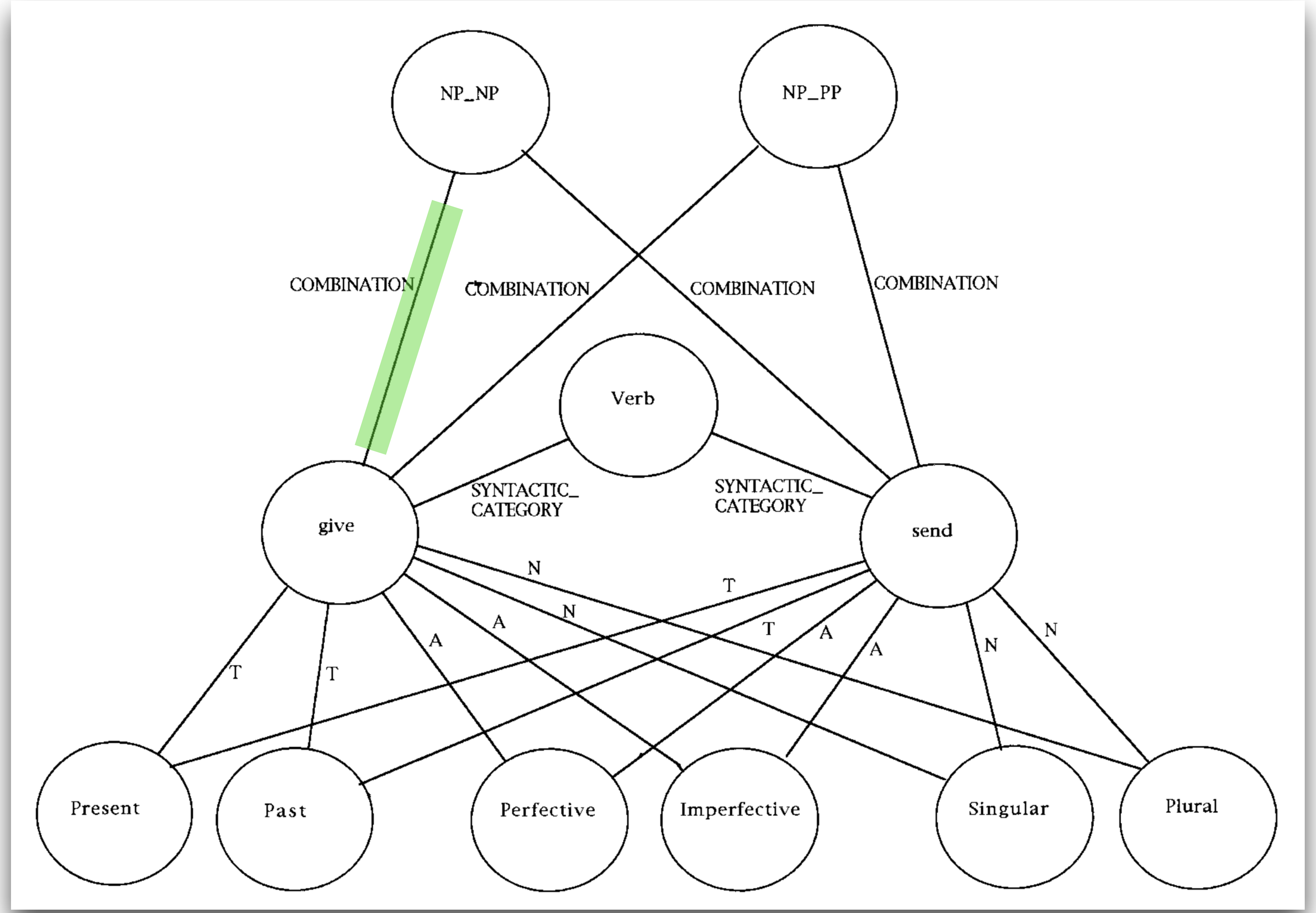


Structural Nodes

Lexical Nodes

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

Theory 2: Implicit Learning



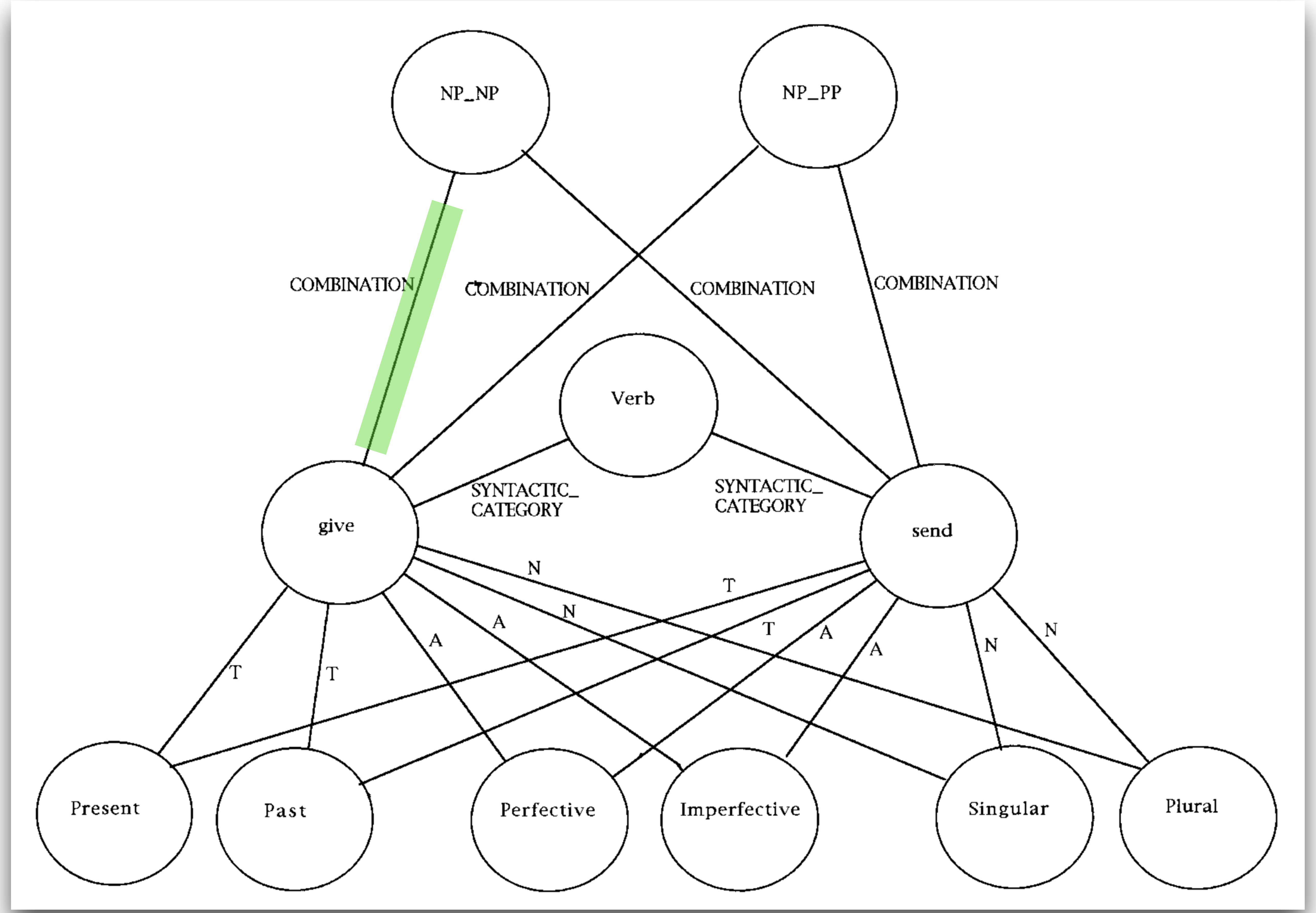
Structural Nodes

Lexical Nodes

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

Theory 2: Implicit Learning

Slowly decaying weight changes



Structural Nodes

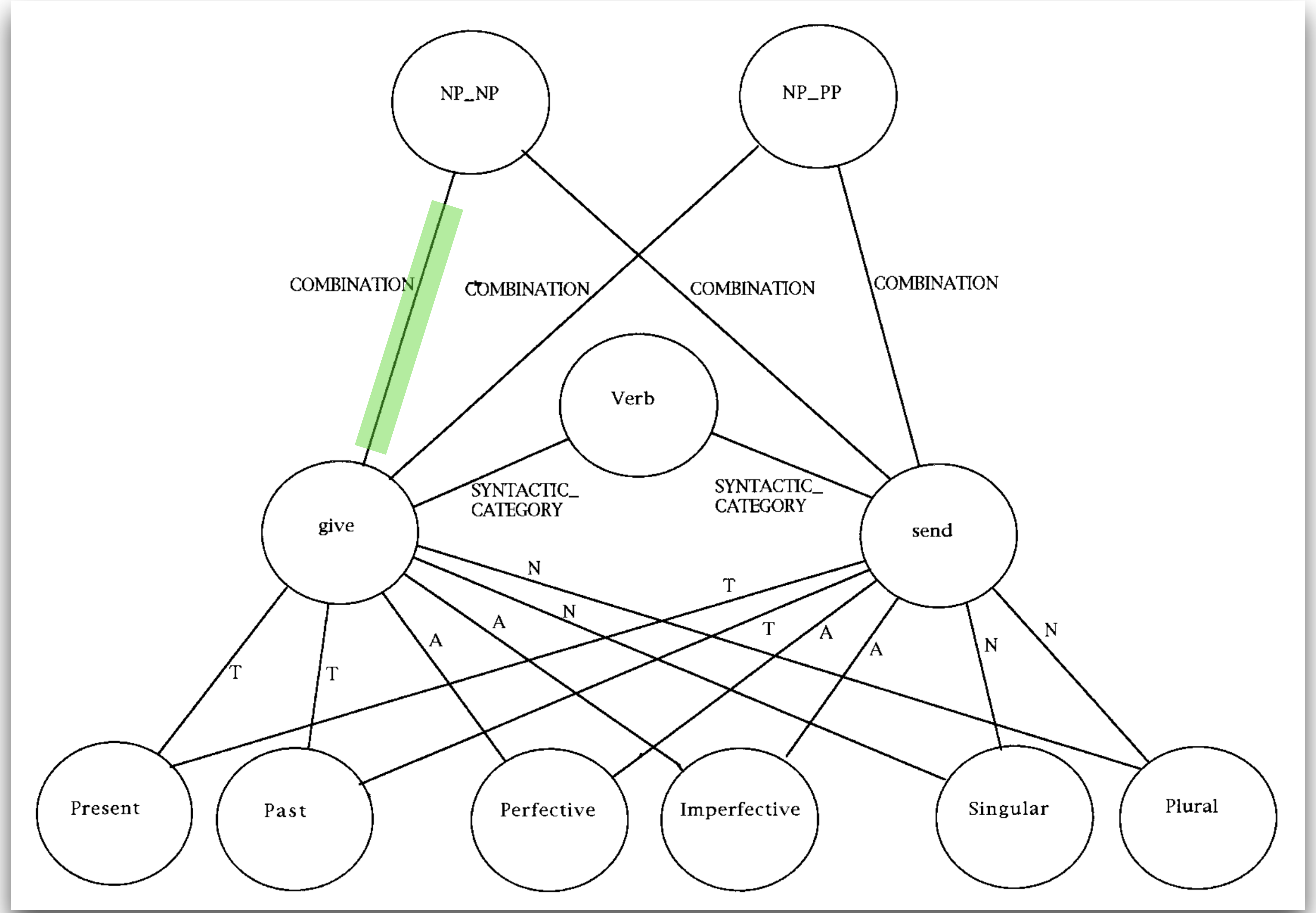
Increasing the link weights <u>proportional to</u> the error signal between verb-bias prediction and the actual input prime.

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.

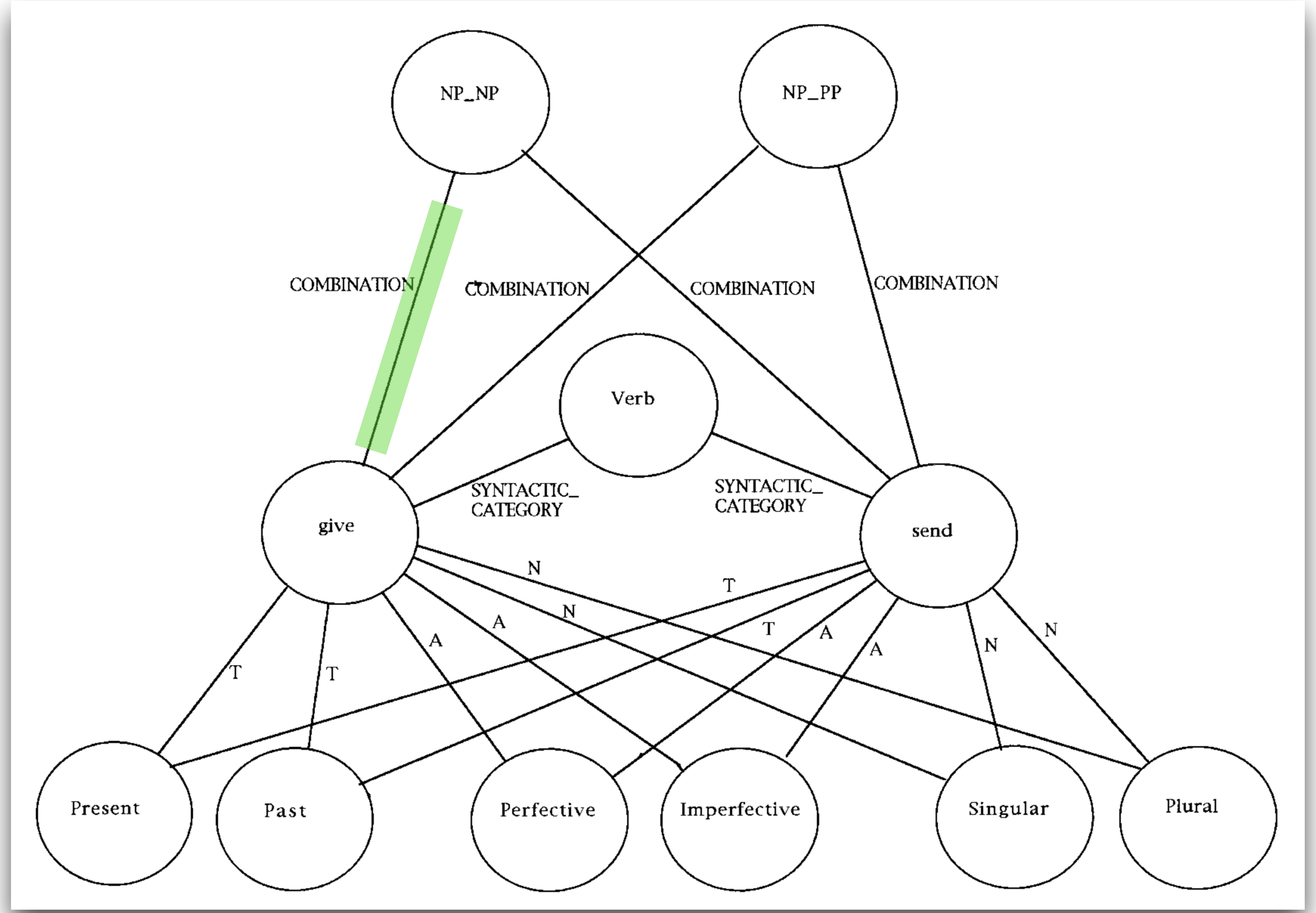
Lexical Nodes

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

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

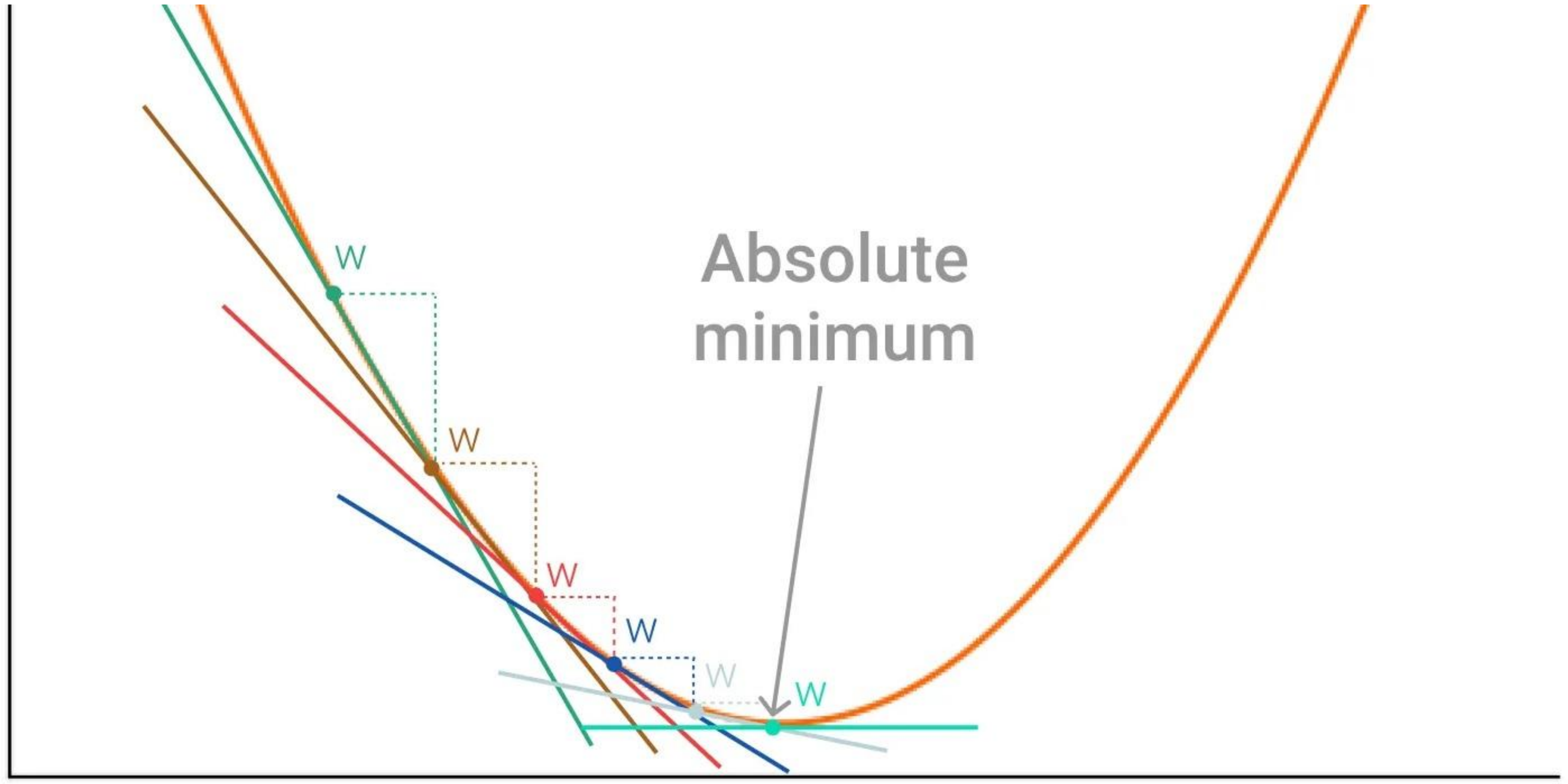
Theory 2: Implicit Learning

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.



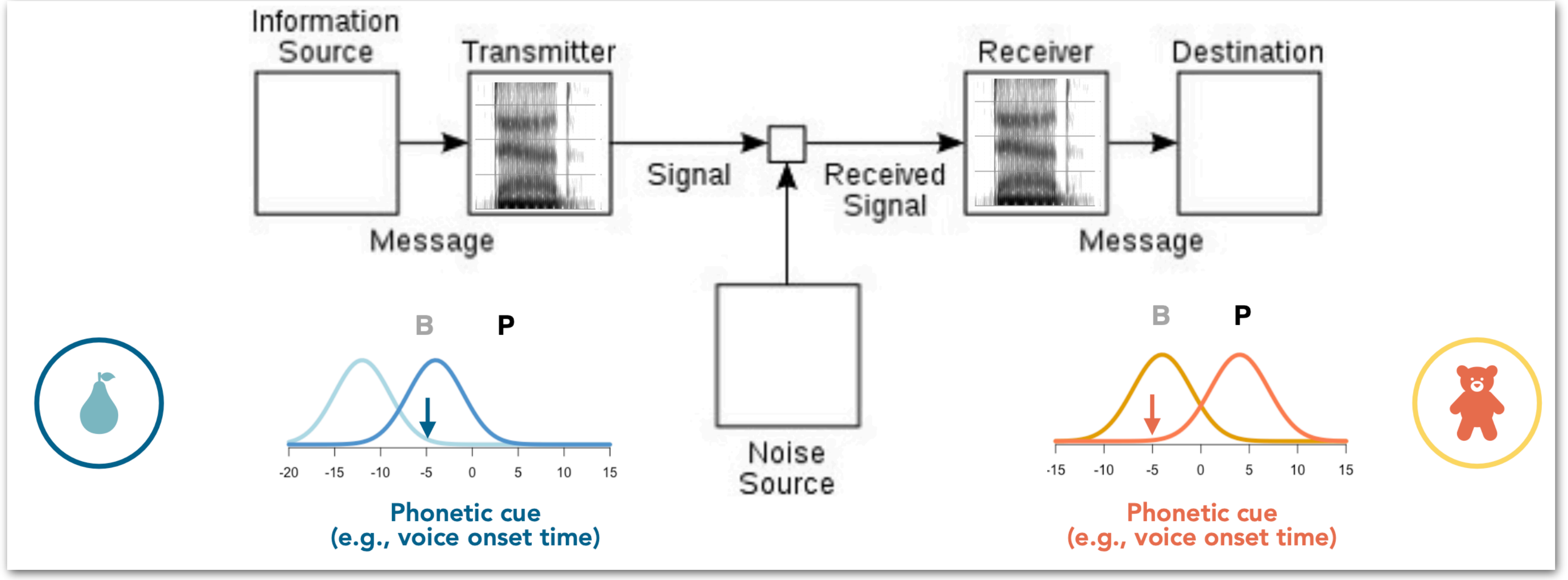
[e.g., Chang et al. 2002, 2006; Jaeger & Snider 2013]

Side Notes: Priming as Linguistic Adaptation

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

Side Notes: Priming as Linguistic Adaptation

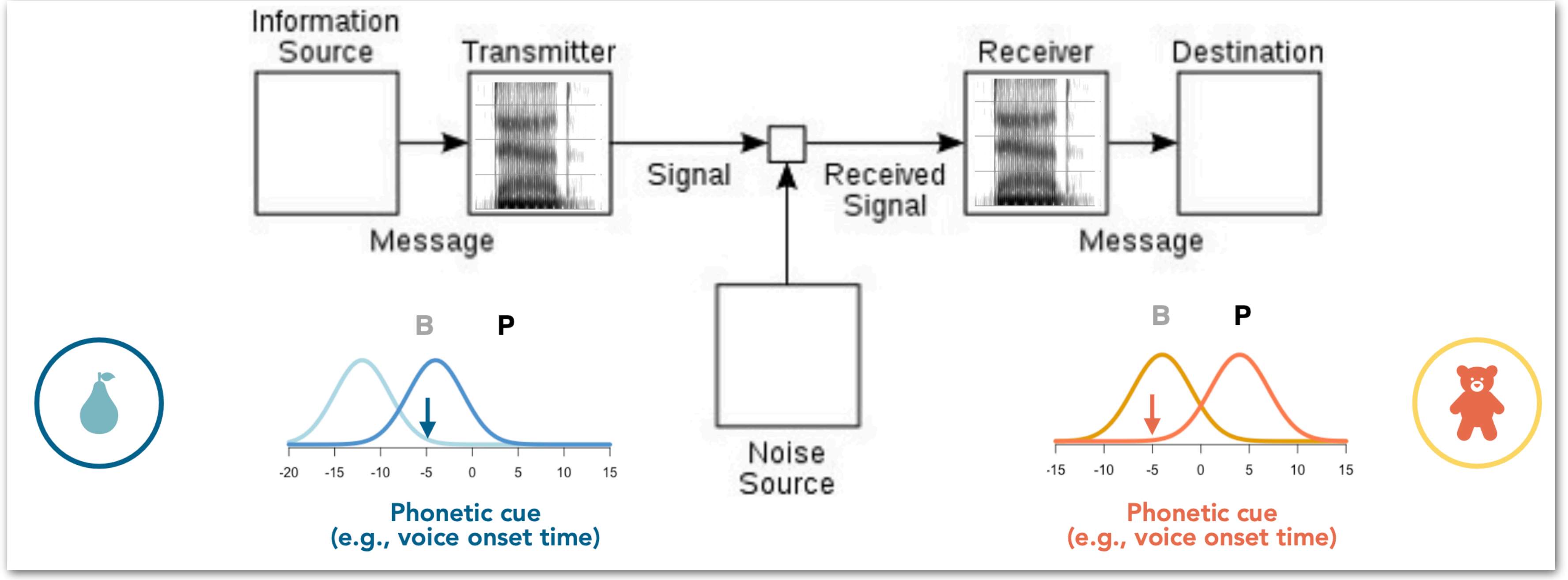
- **Linguistic Adaptation:** adjustment of distributional knowledge / expectations to accommodate local input;
  - Speech adaptation & perception;
  - Syntactic satiation;
  - Informativeness calibration & Presupposition/pragmatic routine calibration;



[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;



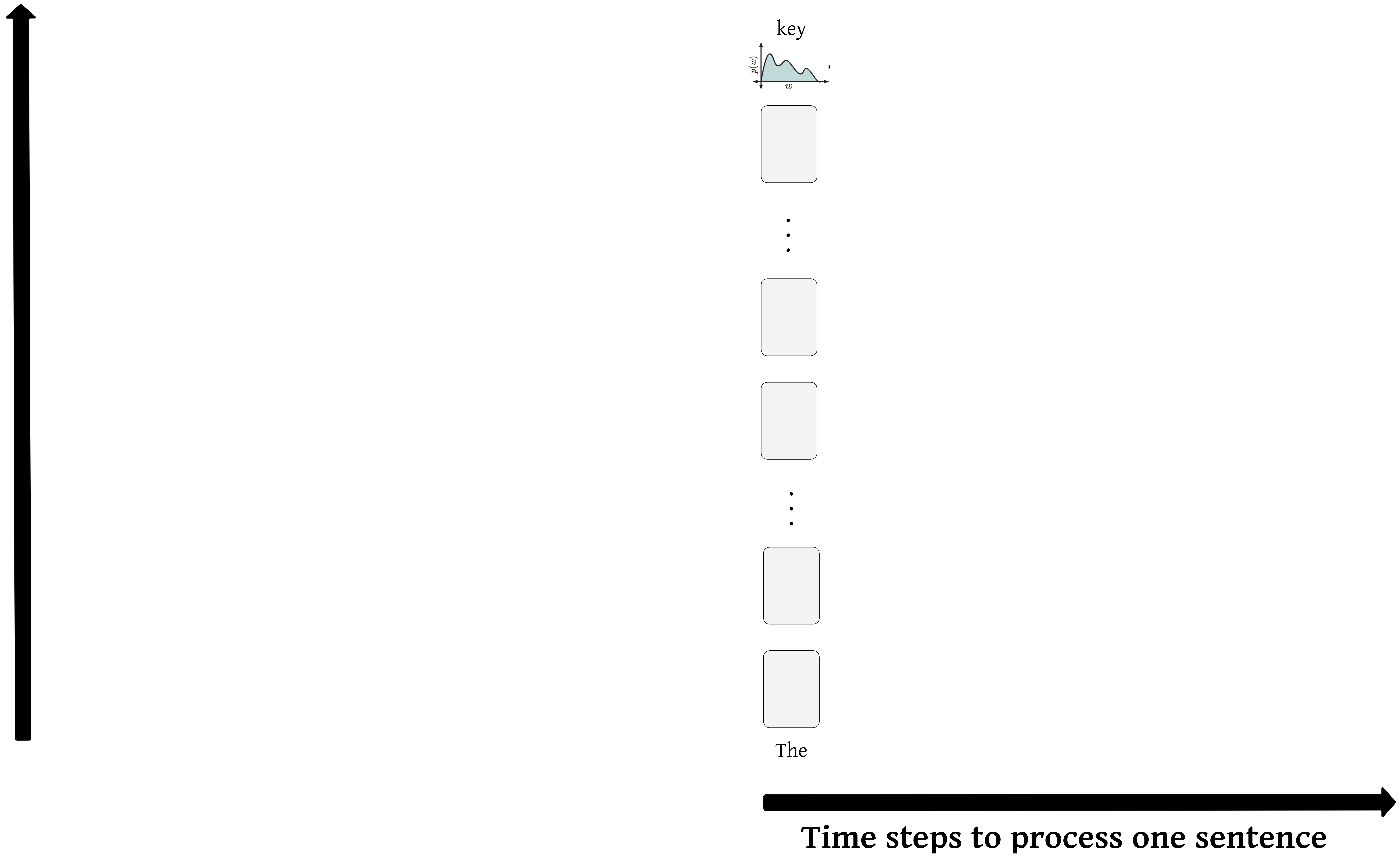
  - Speech adaptation & perception;
  - Syntactic satiation;
  - Informativeness calibration & Presupposition/pragmatic routine calibration;

- **Linguistic Adaptation & Statistical Learning?**
  - Adaptation as a *micro-learning window*: you can watch the learner update after tiny bits of evidence.
  - **IFE is the *signature* that the update is error-driven**, not just residual activation: rarer structures create bigger prediction error → bigger update.

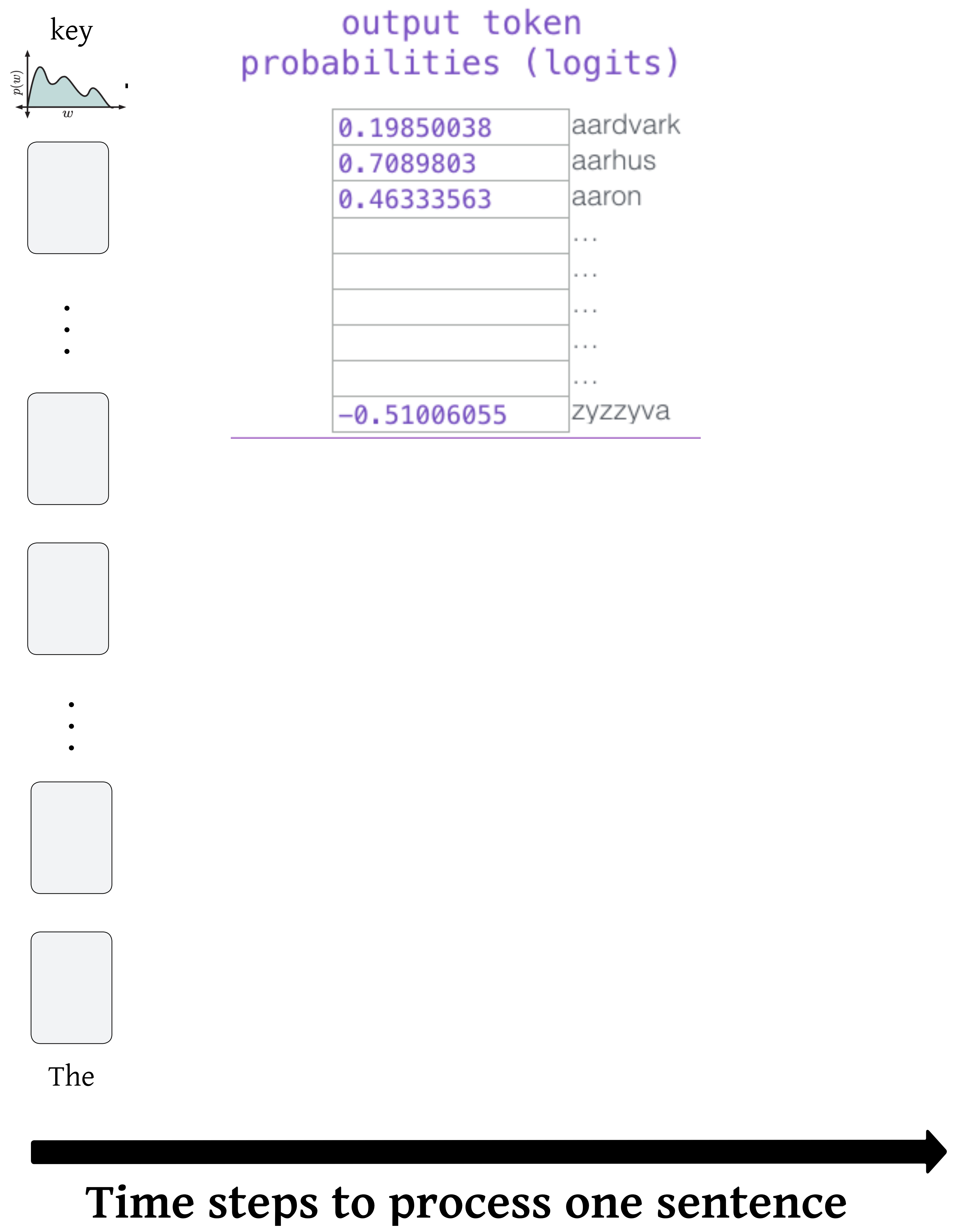
[e.g., Xie & Kurumada 2023; Lu et al. 2025; Feng 2022; etc.]

In-Context Learning in LLMs

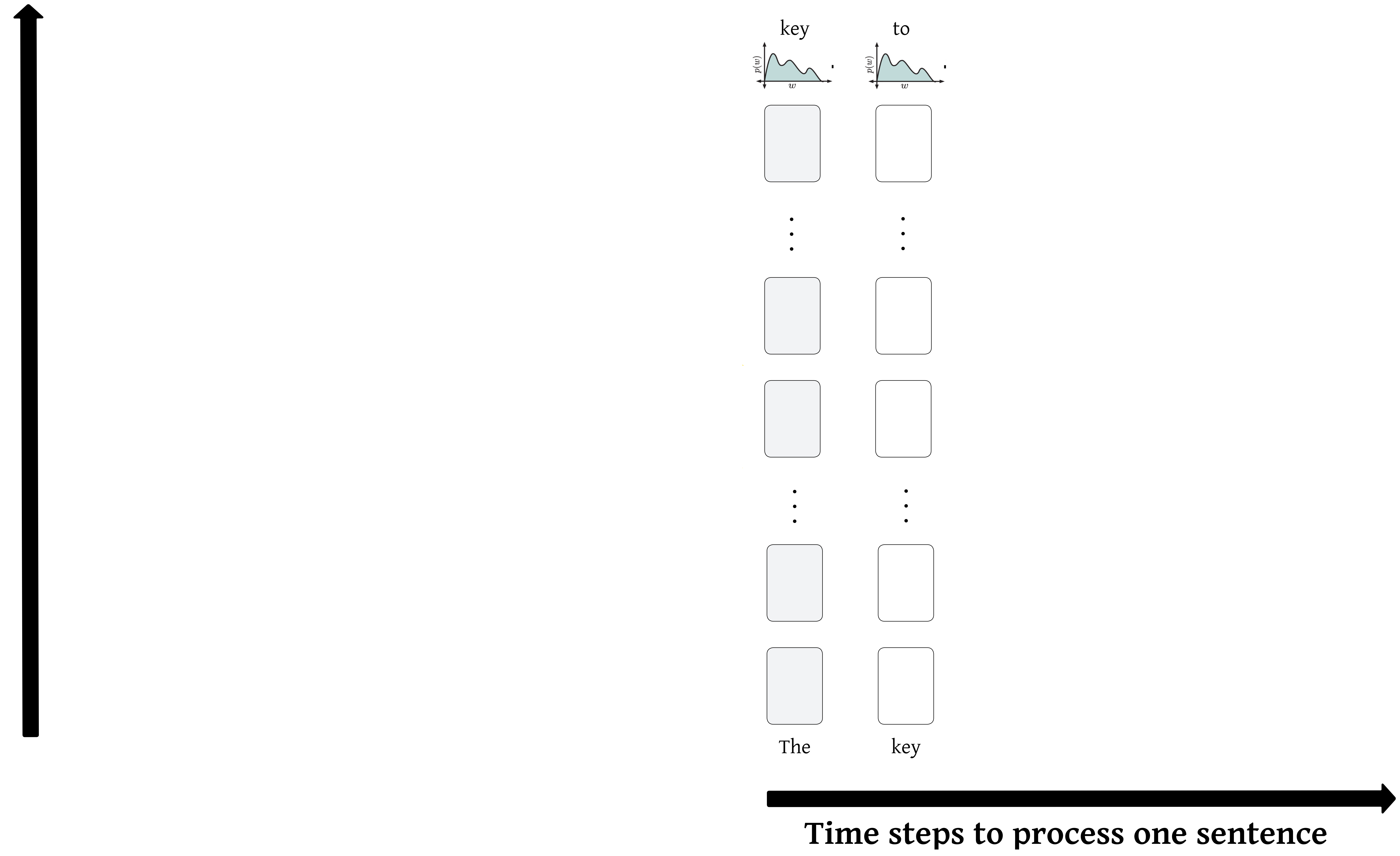
A Primer on Intuitions behind LLMs



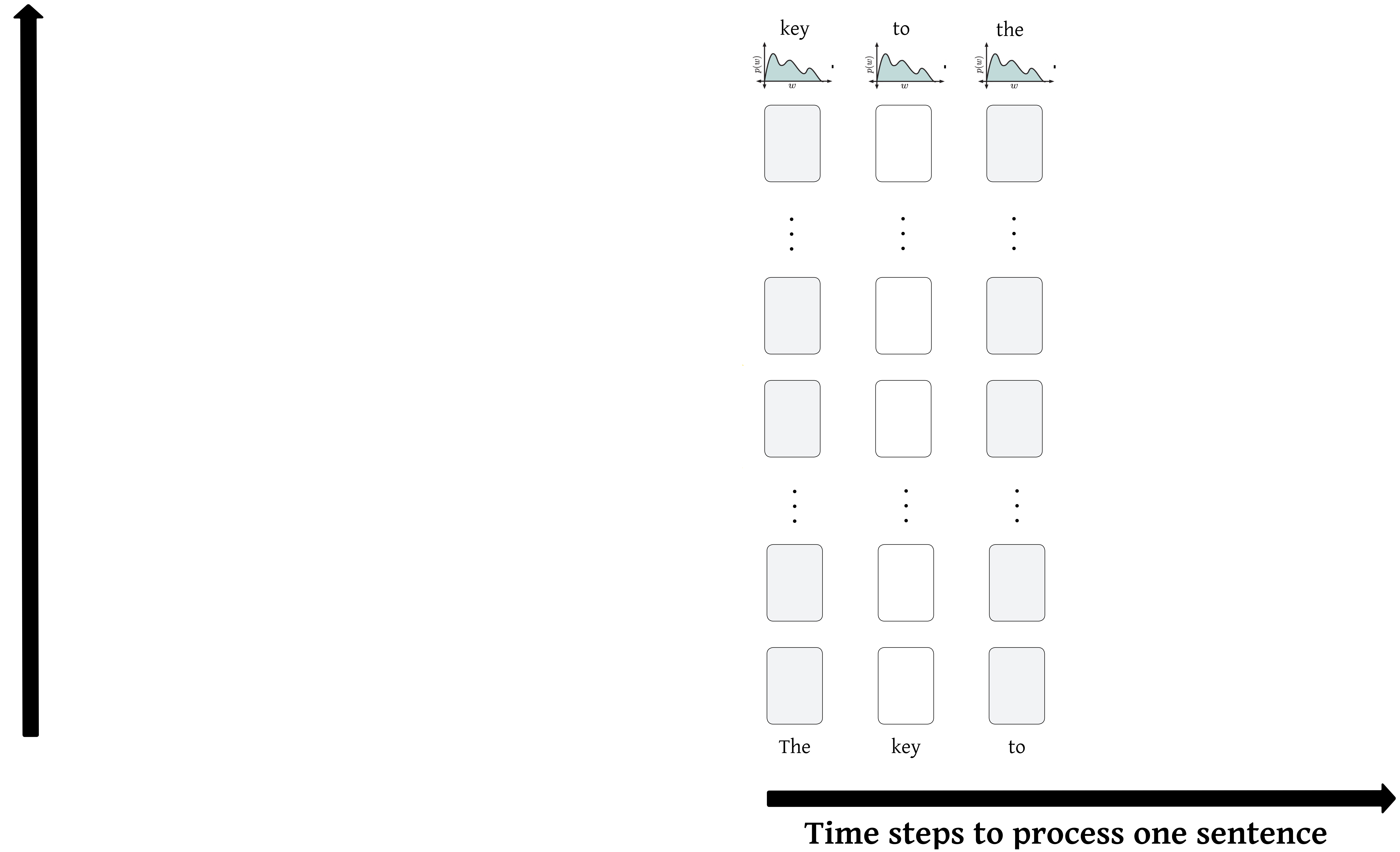
A Primer on Intuitions behind LLMs



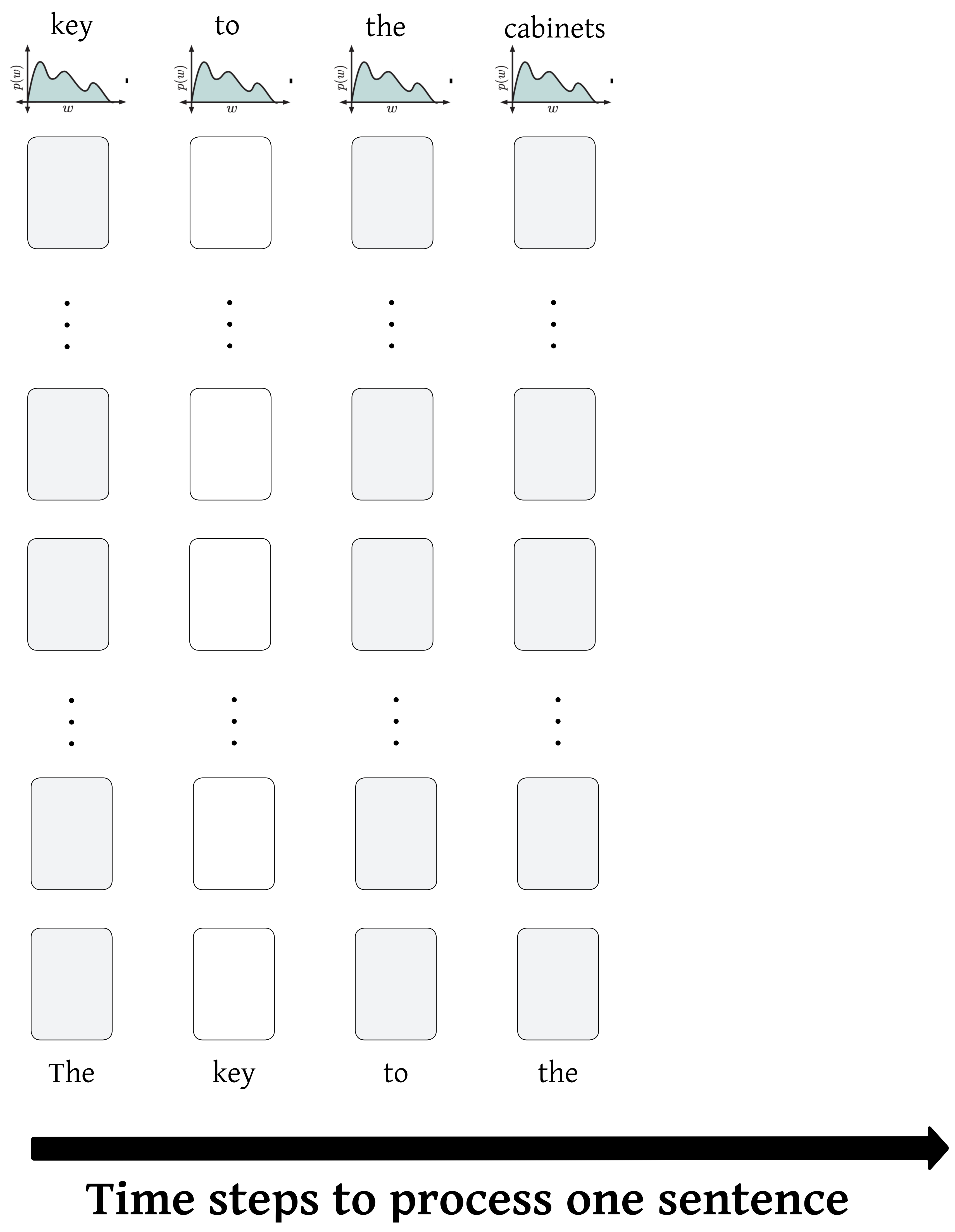
A Primer on Intuitions behind LLMs



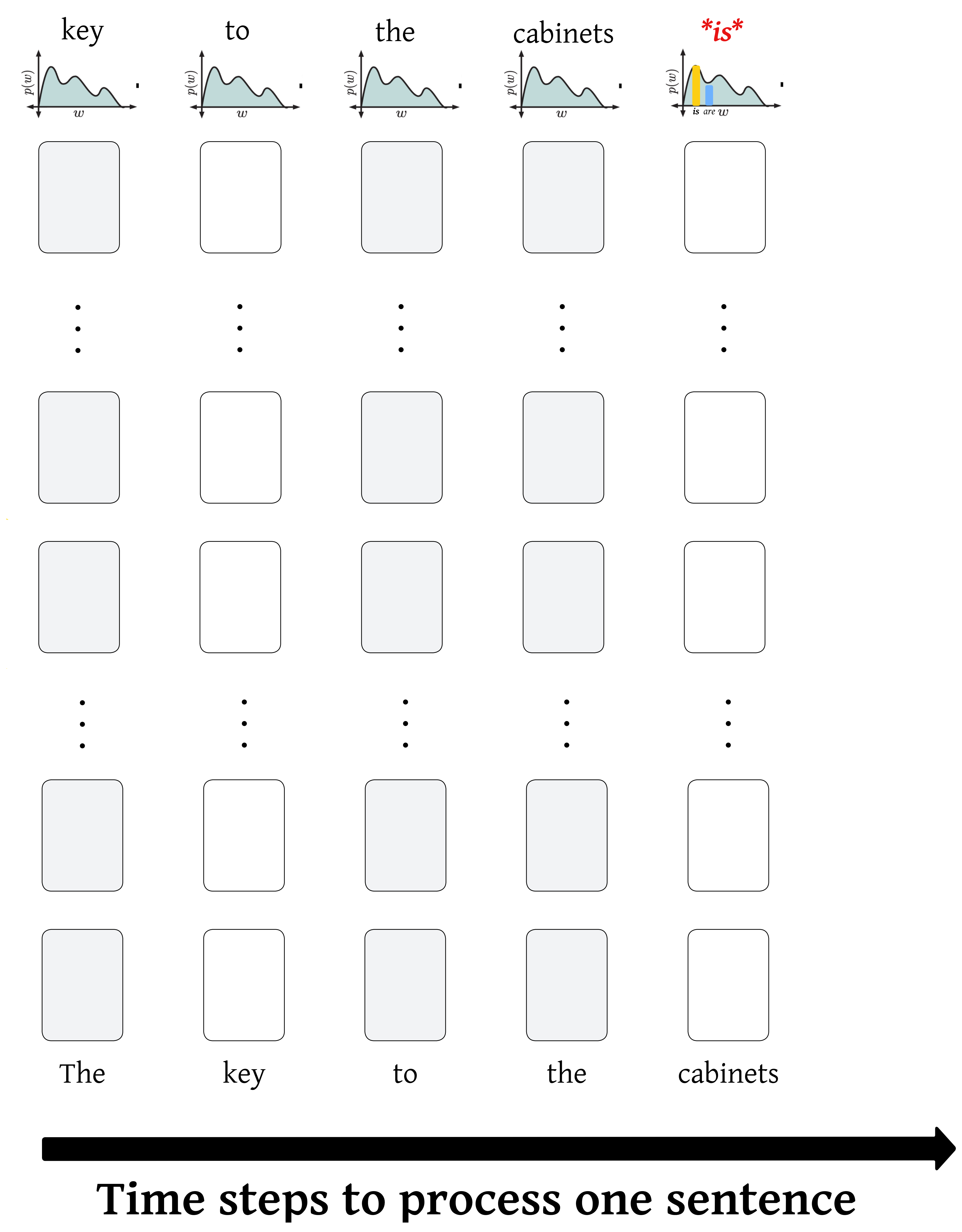
A Primer on Intuitions behind LLMs



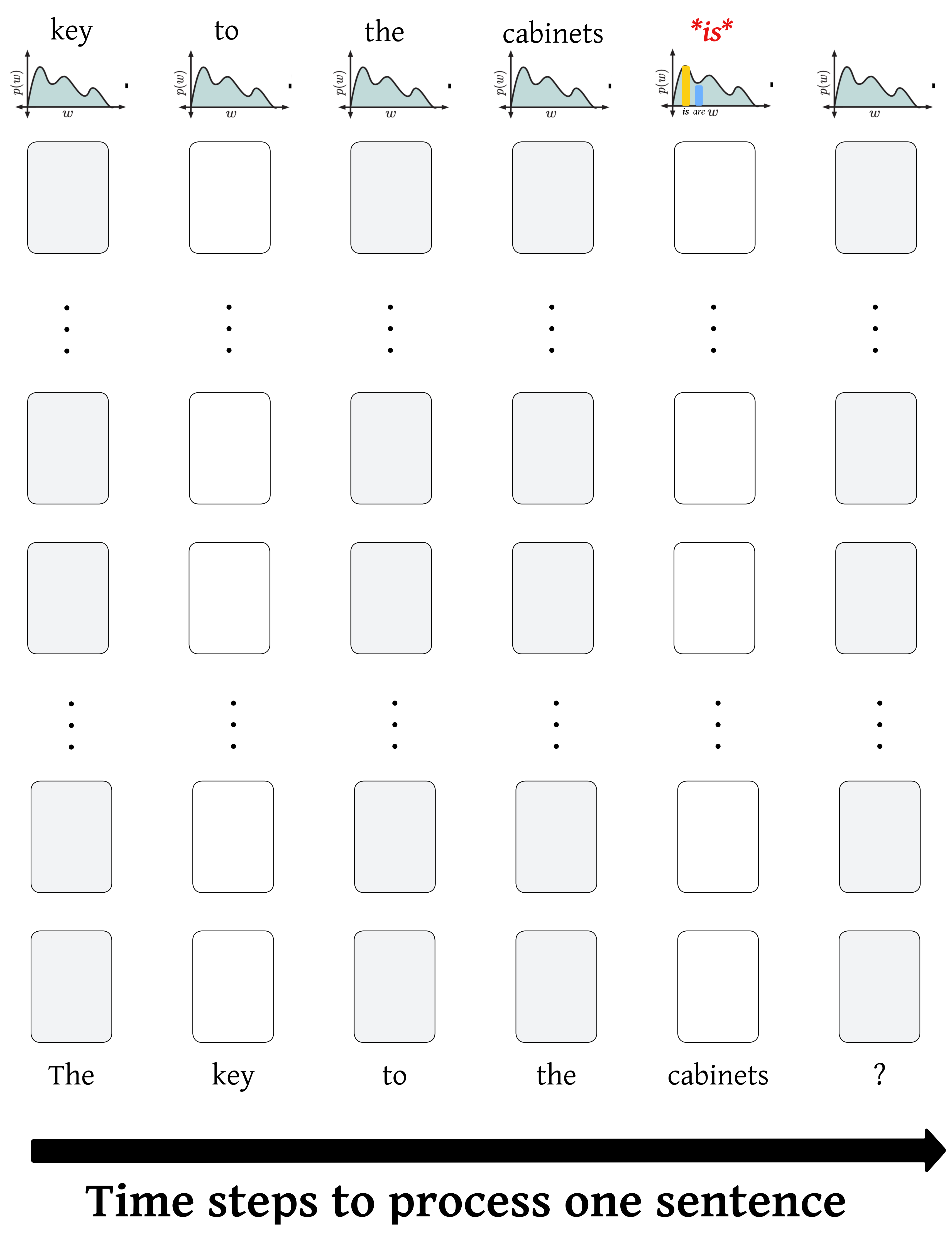
A Primer on Intuitions behind LLMs



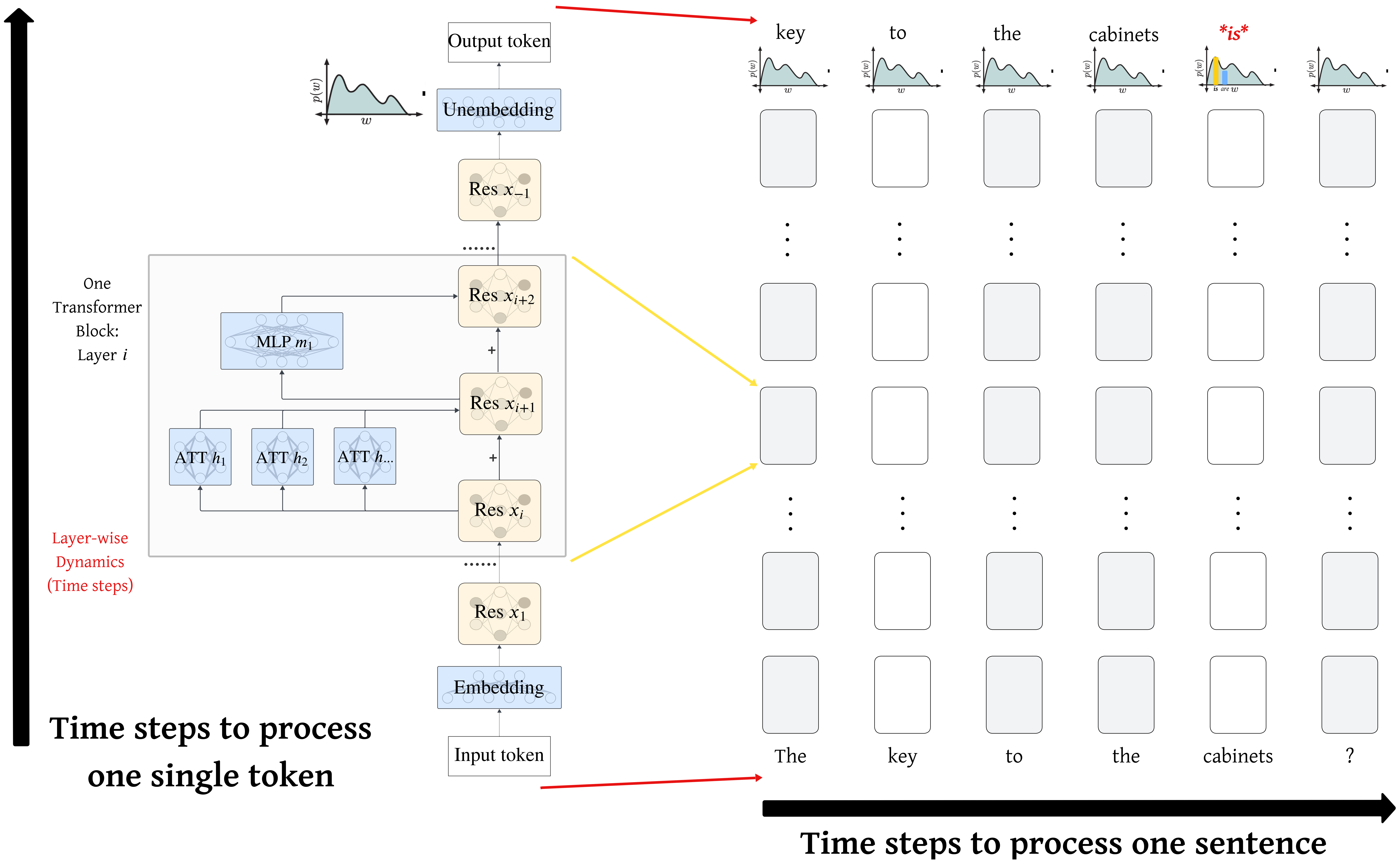
A Primer on Intuitions behind LLMs



A Primer on Intuitions behind LLMs



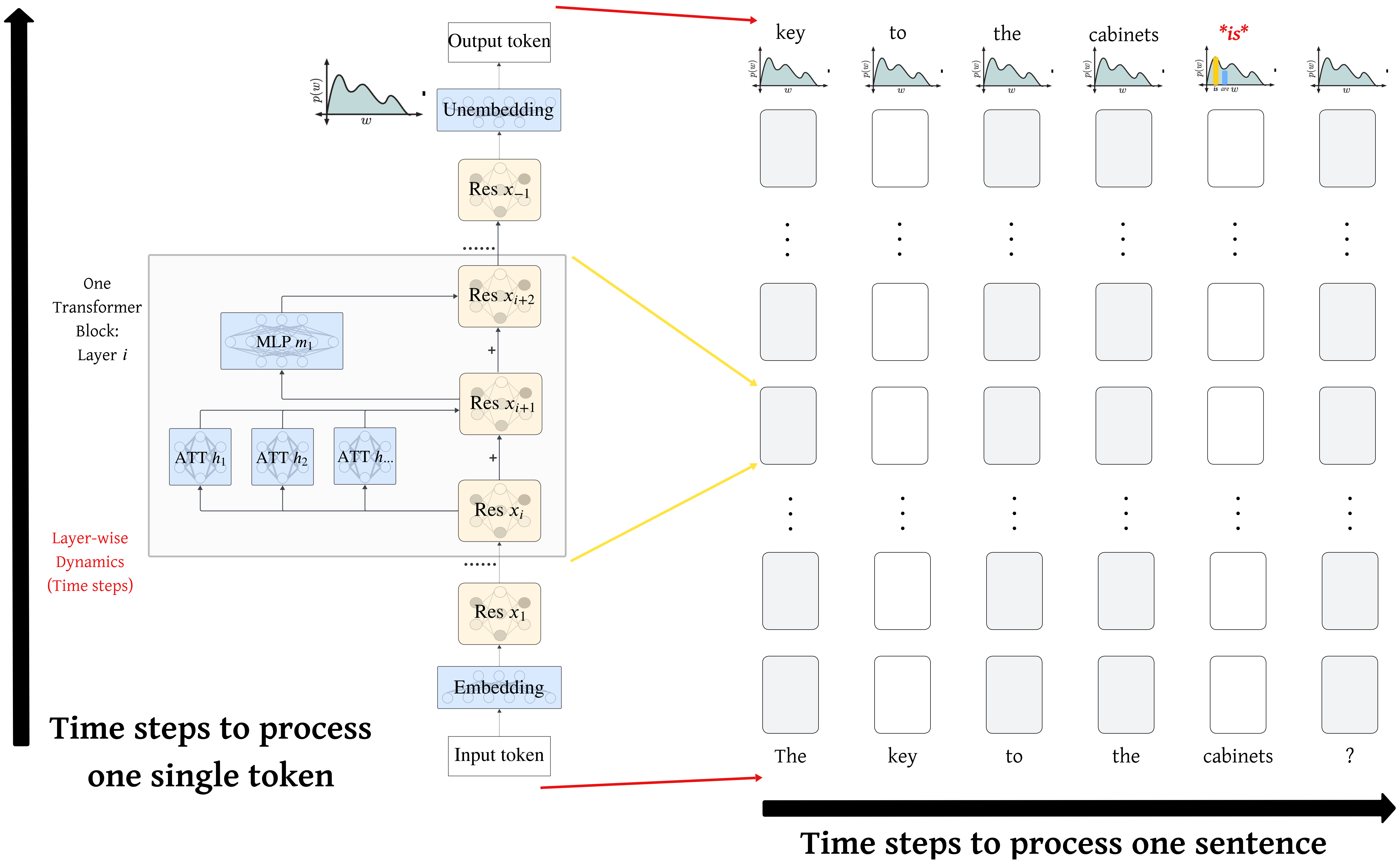
A Primer on Intuitions behind LLMs



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!

1 ==> 4
2 ==> 8
5 ==> 20
8 ==> ??

[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!

&() --> ()&

234 --> 342

@#$ --> #$@

rte--->

[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!

1 * 1 = 2
3 * 3 = 6
4 * 2 = 6
7 * 1 = ??

[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

$1 * 1 = 2$

$3 * 3 = 6$

$4 * 2 = 6$

$7 * 1 = ??$

[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.

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

Few-shot

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

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

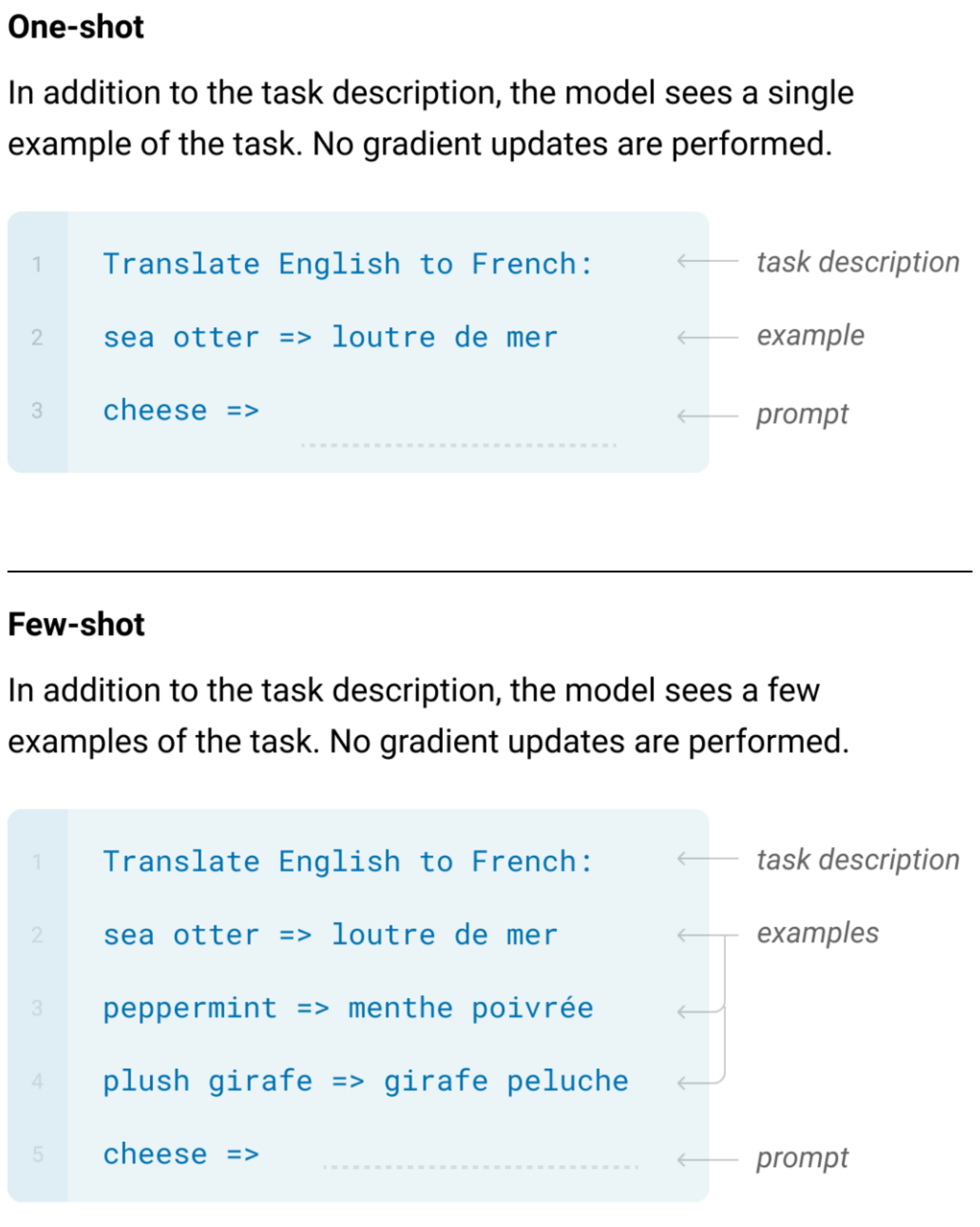
[Brown et al. 2020]

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



In-context Learning:

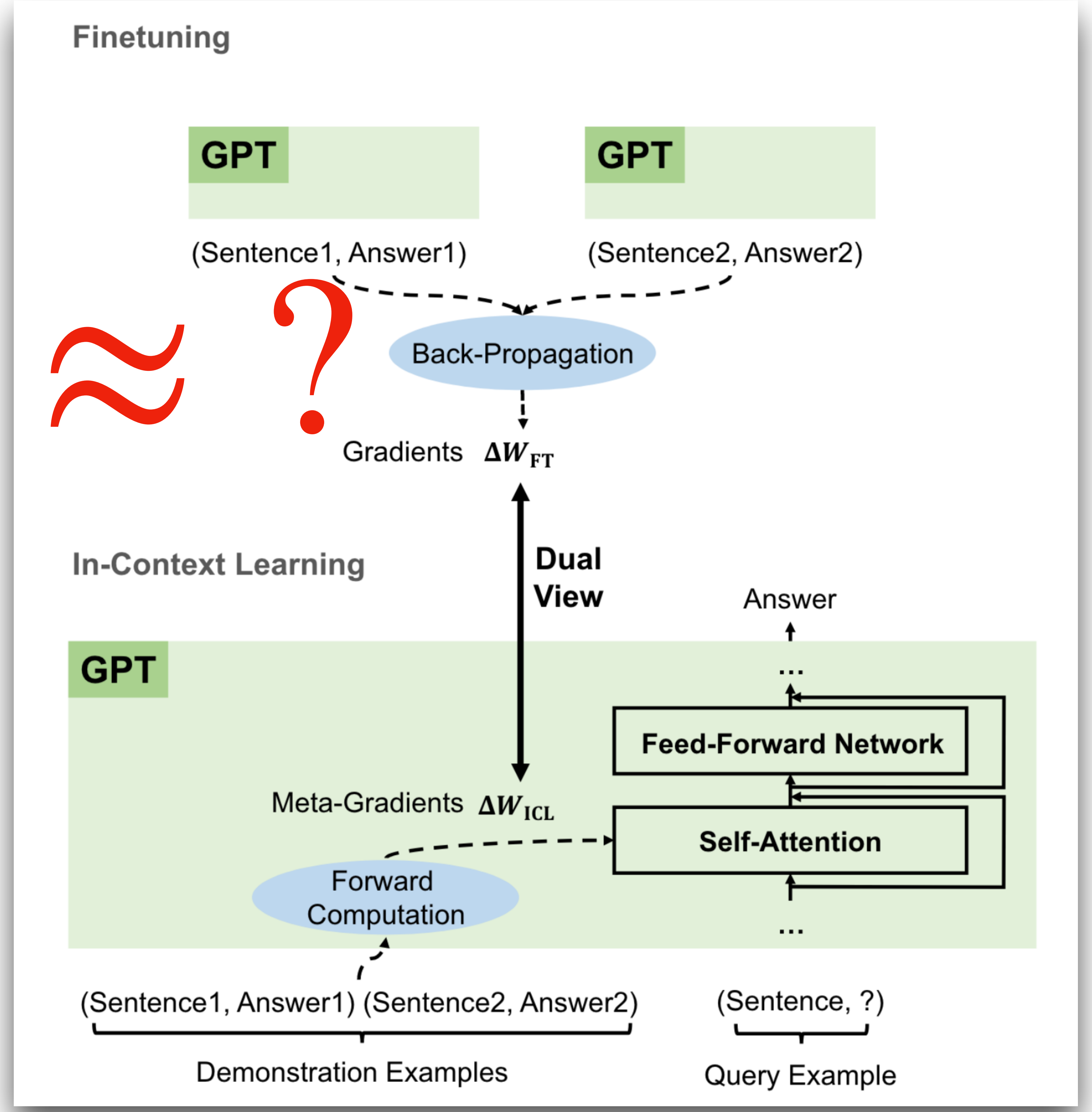
- No gradient updates;
- Rapid: from a few examples;
- One-shot / Few-shot learning;

In-weights Learning (Fine-tuning):

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

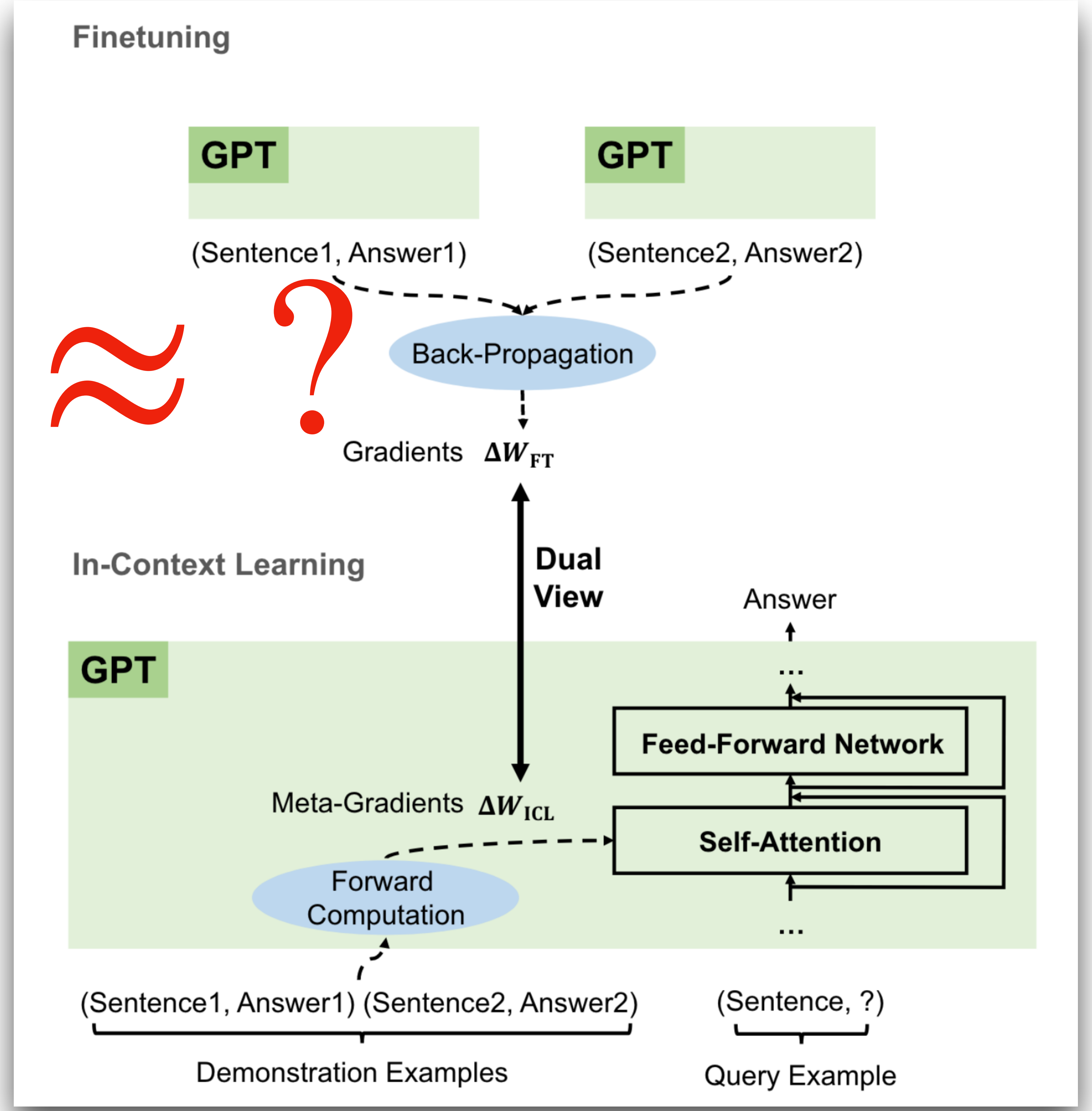
[Brown et al. 2020]

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?

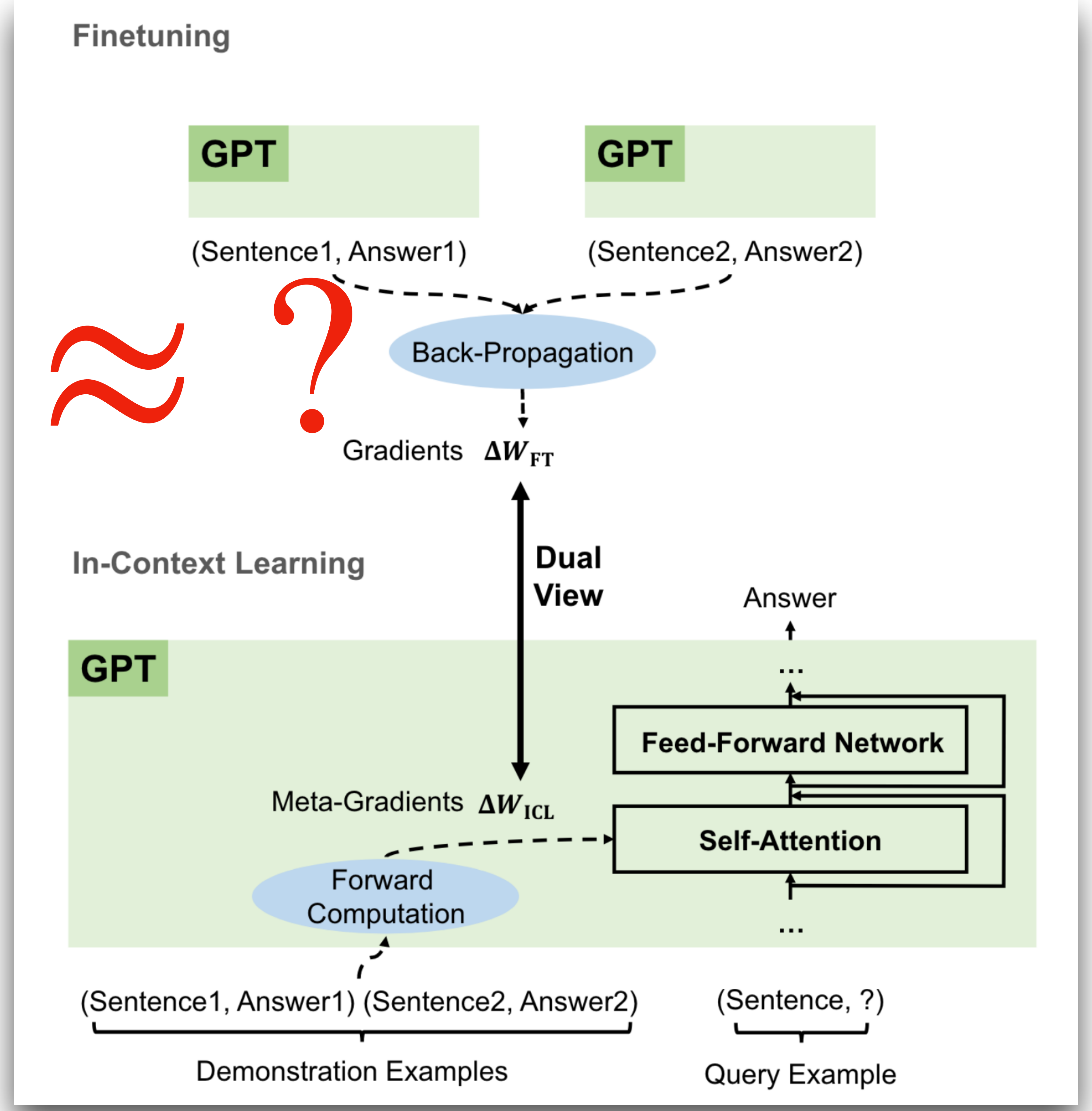


Previous studies in AI/CS:

- On toy models and toy tasks;
- In-principle claims: math derivations;

[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 studies in AI/CS:

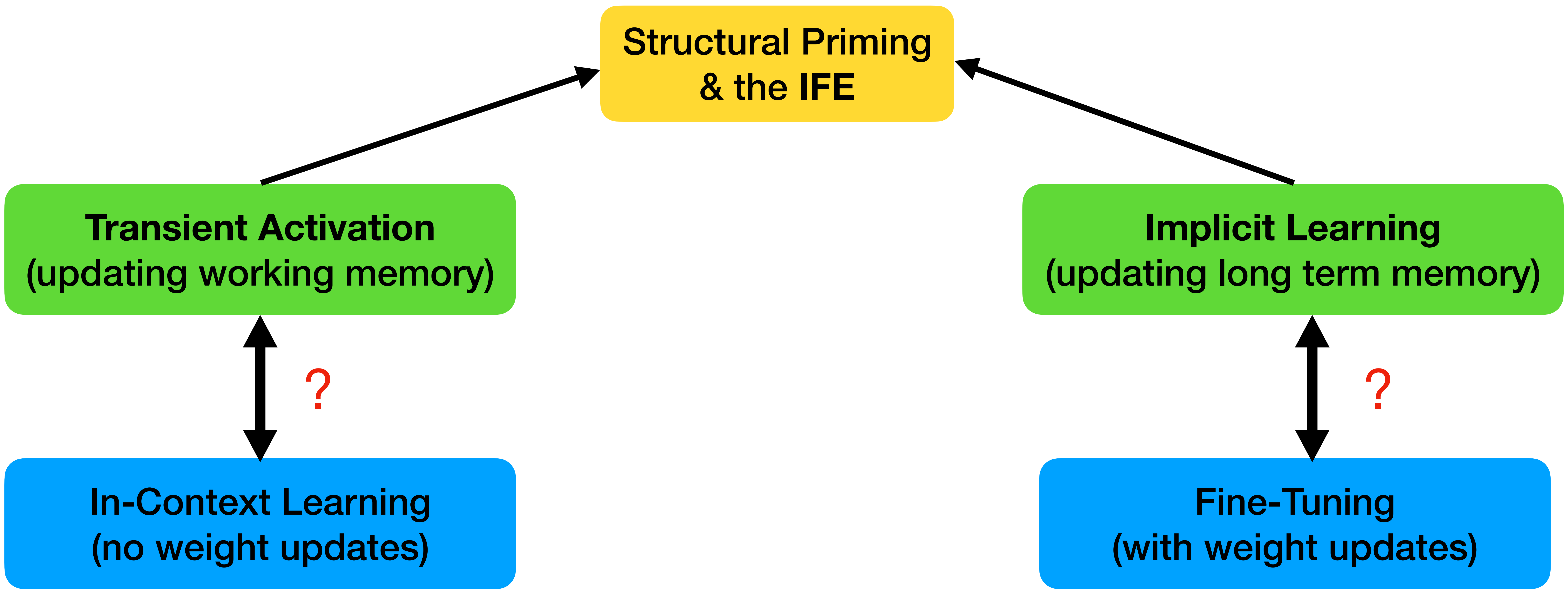
- On toy models and toy tasks;
- In-principle claims: math derivations;

Current work:

- On pretrained, actual LLMs;
- Psycholinguistically motivated — *after all, LLMs are trained to model language* 🙂

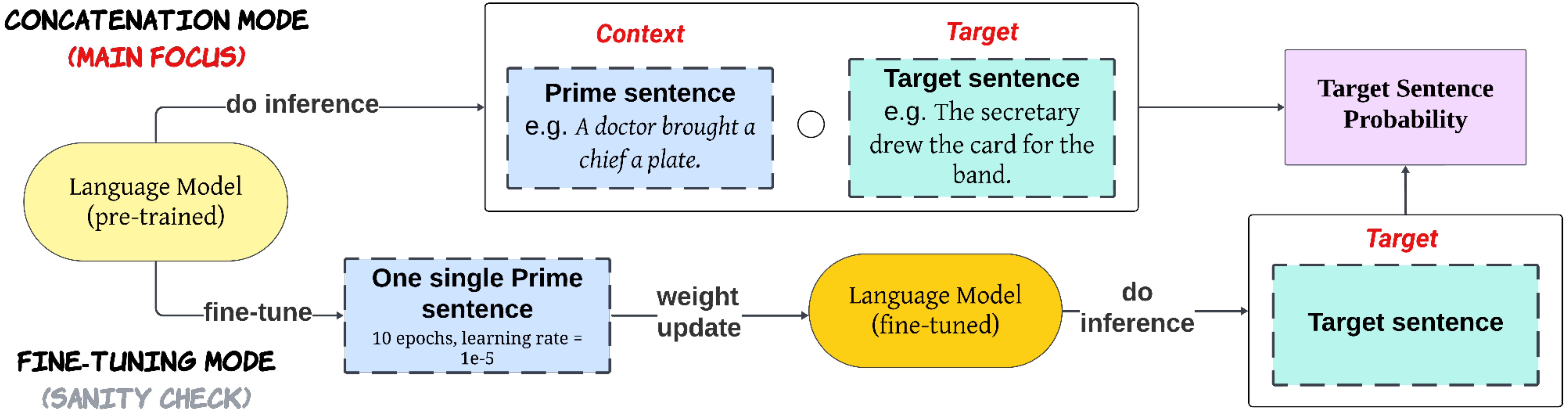
[e.g., Xie et al. 2022, von Oswald et al. 2022, Dai et al. 2023]

Current Picture



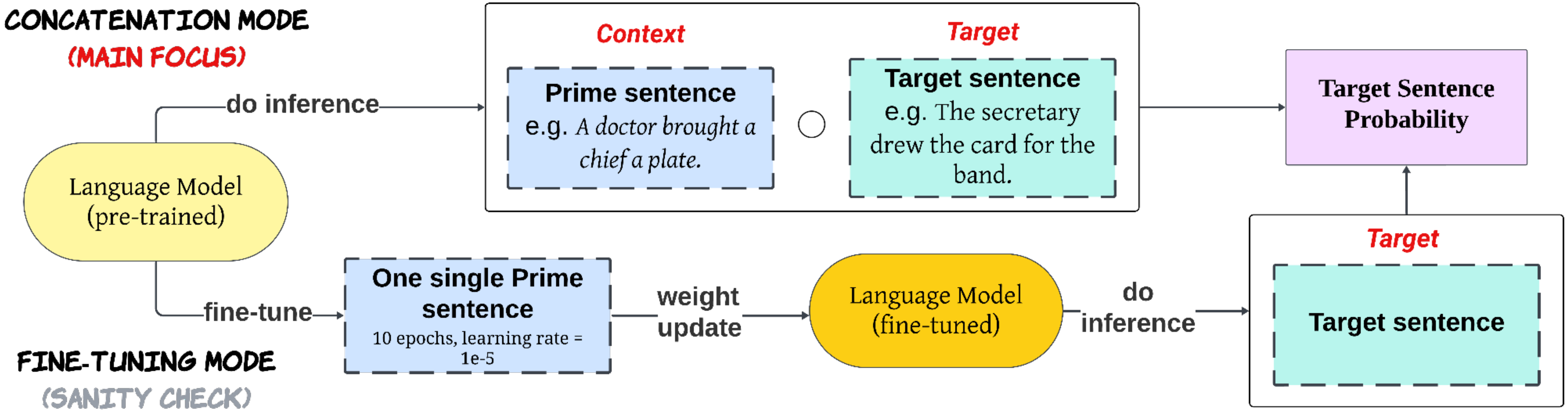
Experiment 1: Behavioral Analysis

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

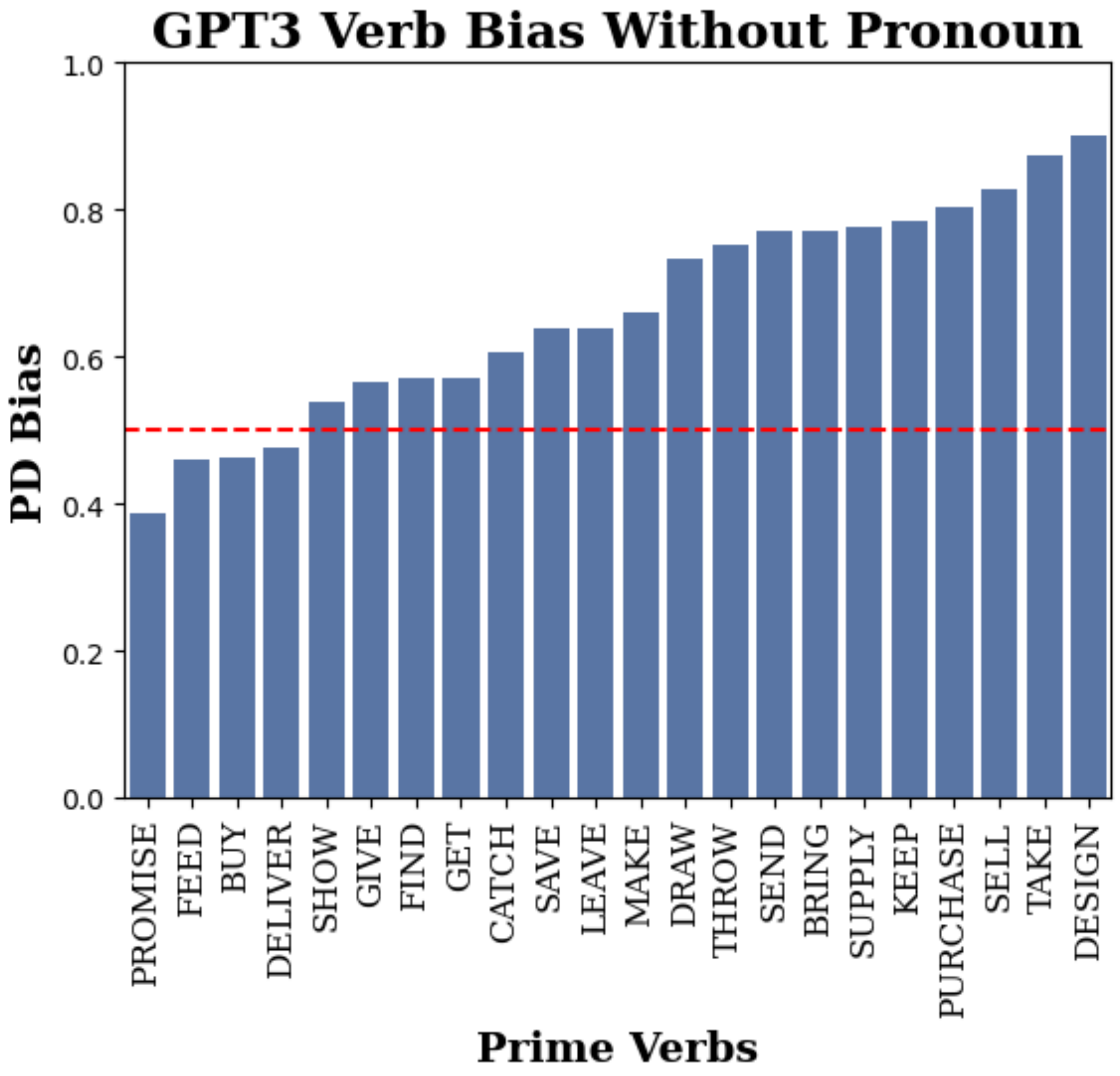
Verb Bias of verb V on structure X:

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

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:

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

Quantifying Verb Biases and the IFE

Verb Bias of verb V on structure X:

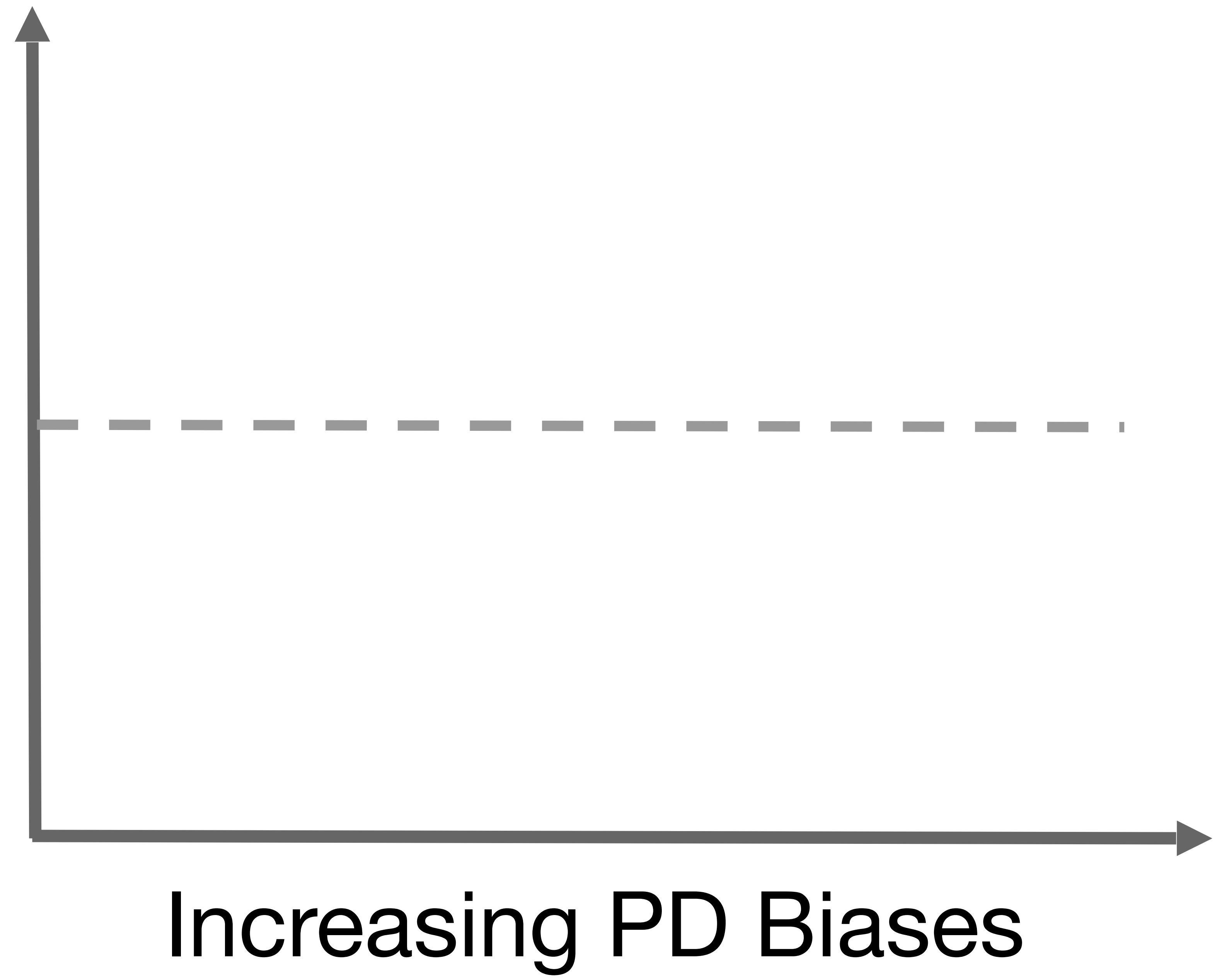
$$bias(V, \mathrm{PD}) = \frac{1}{|\mathcal{S}_V|} \sum_{t_{\mathrm{PD}} \in \mathcal{S}_V} \frac{\mathcal{P}(t_{\mathrm{PD}})}{\mathcal{P}(t_{\mathrm{PD}}) + \mathcal{P}(t_{\mathrm{DO}})}$$

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

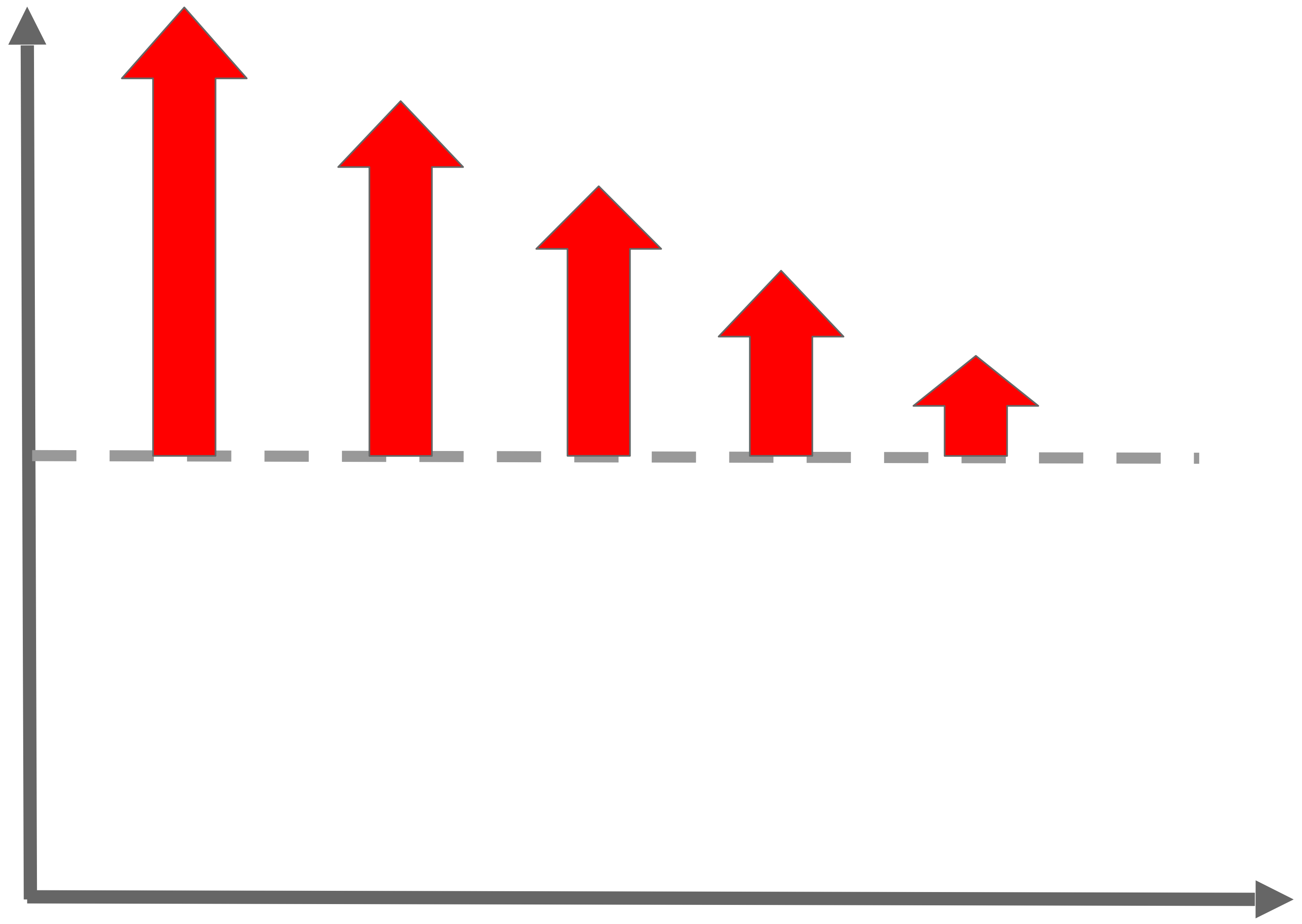
$\text{Prime}_{PD} \rightarrow \text{Target}_{PD}$



Increasing PD Biases

Predictions on the IFE

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

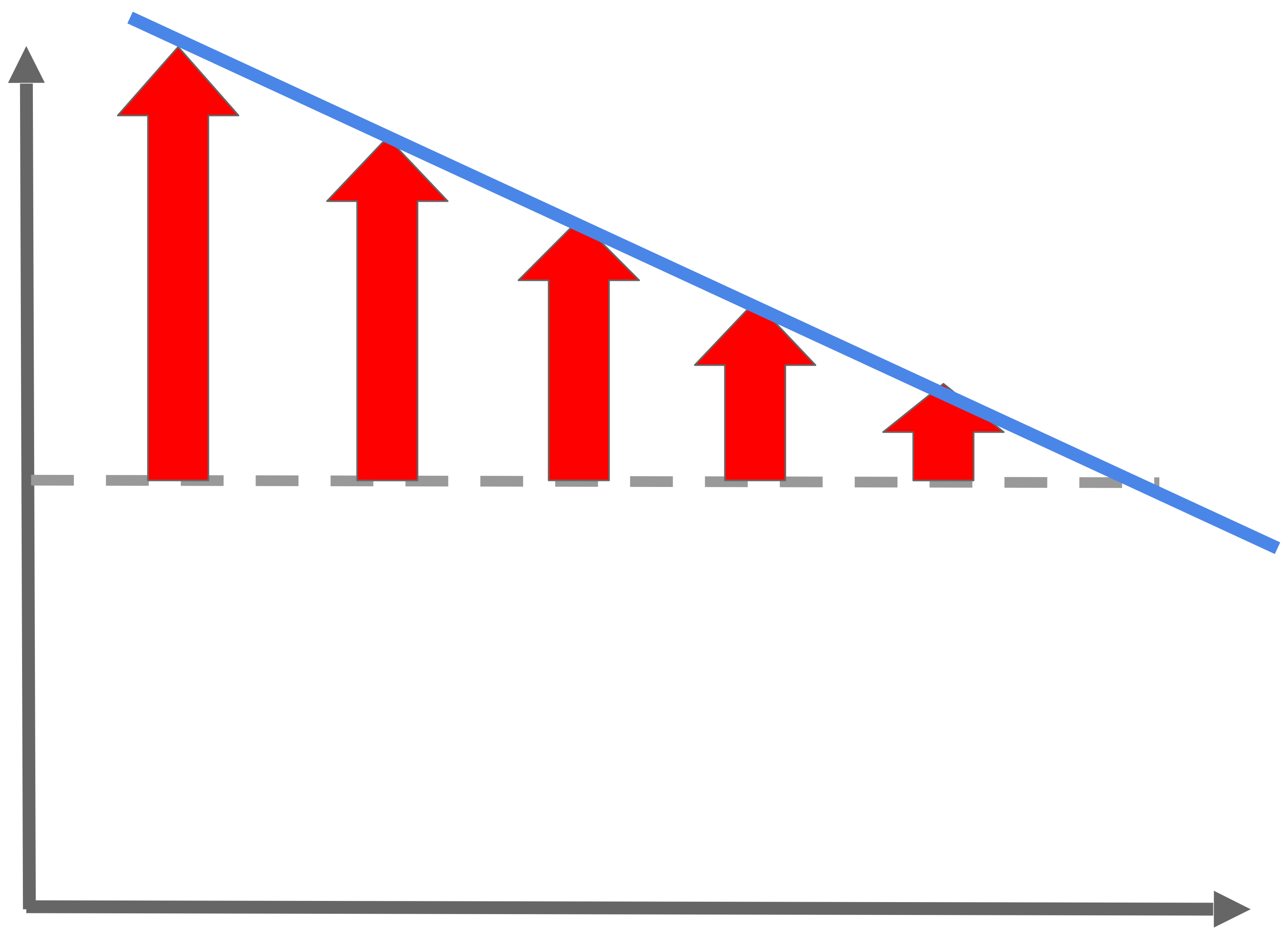


Increasing PD Biases

Increasing PD Biases

Predictions on the IFE

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

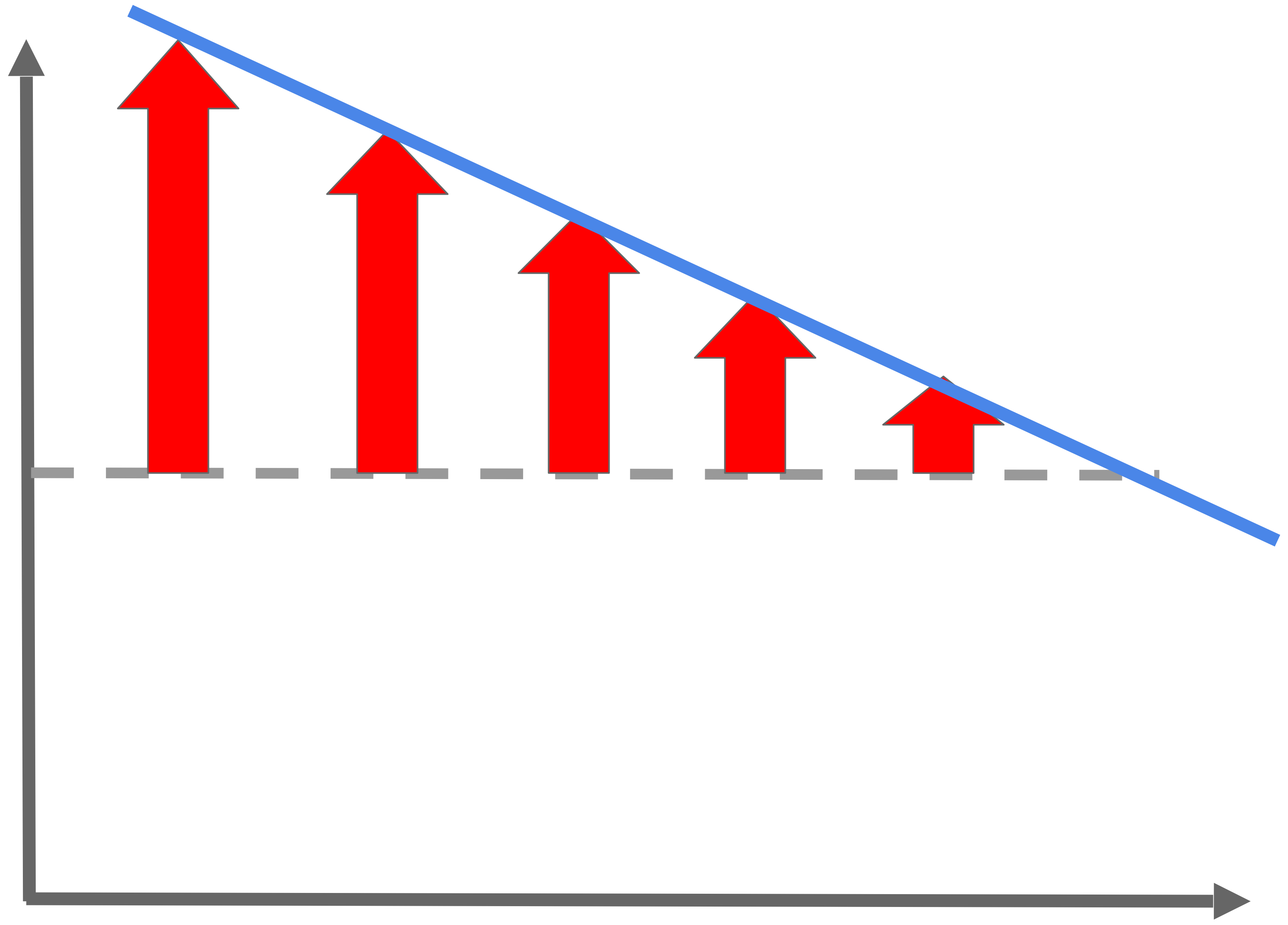


Increasing PD Biases

Increasing PD Biases

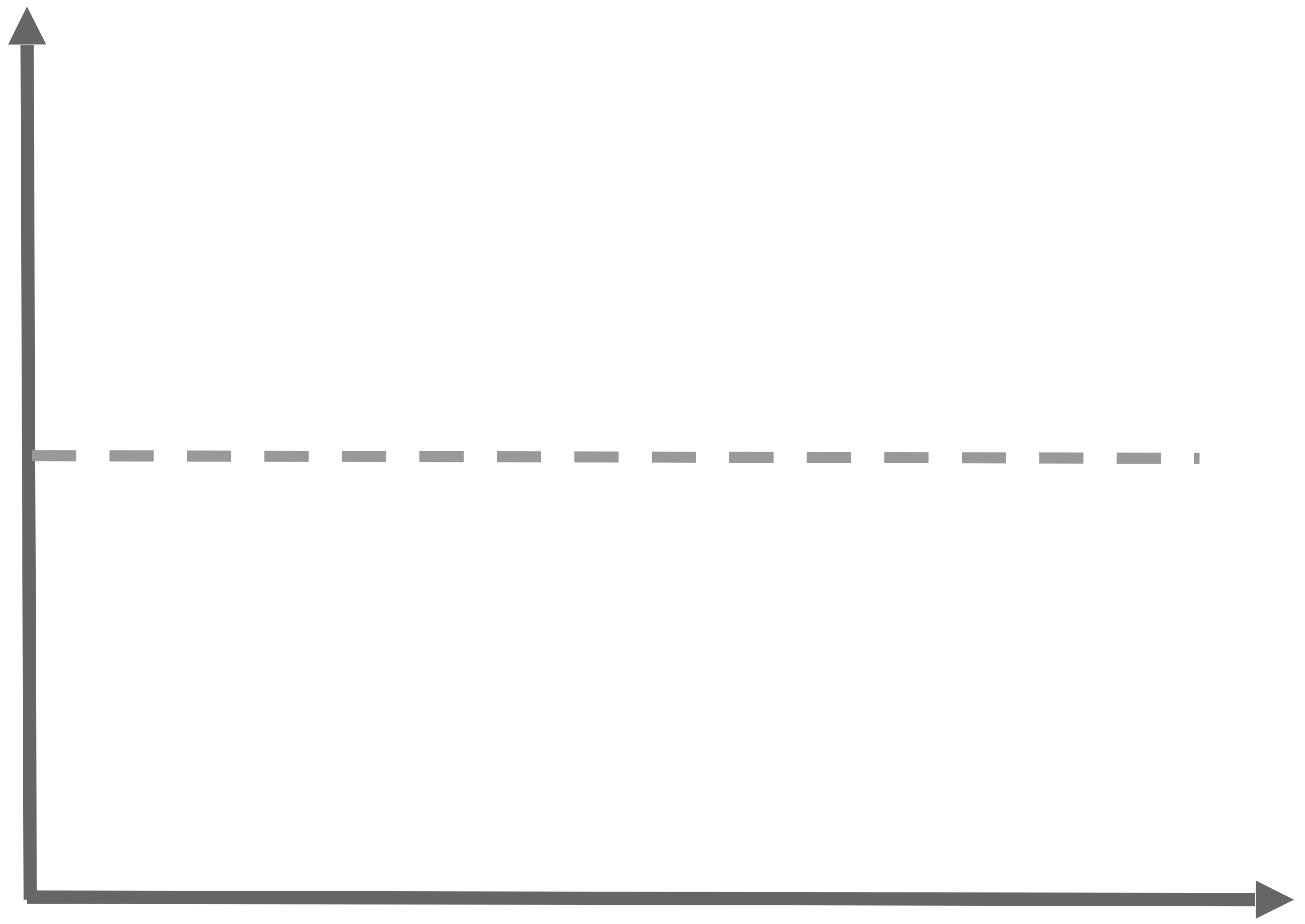
Predictions on the IFE

Prime$_{PD}$ → Target$_{PD}$



Increasing PD Biases

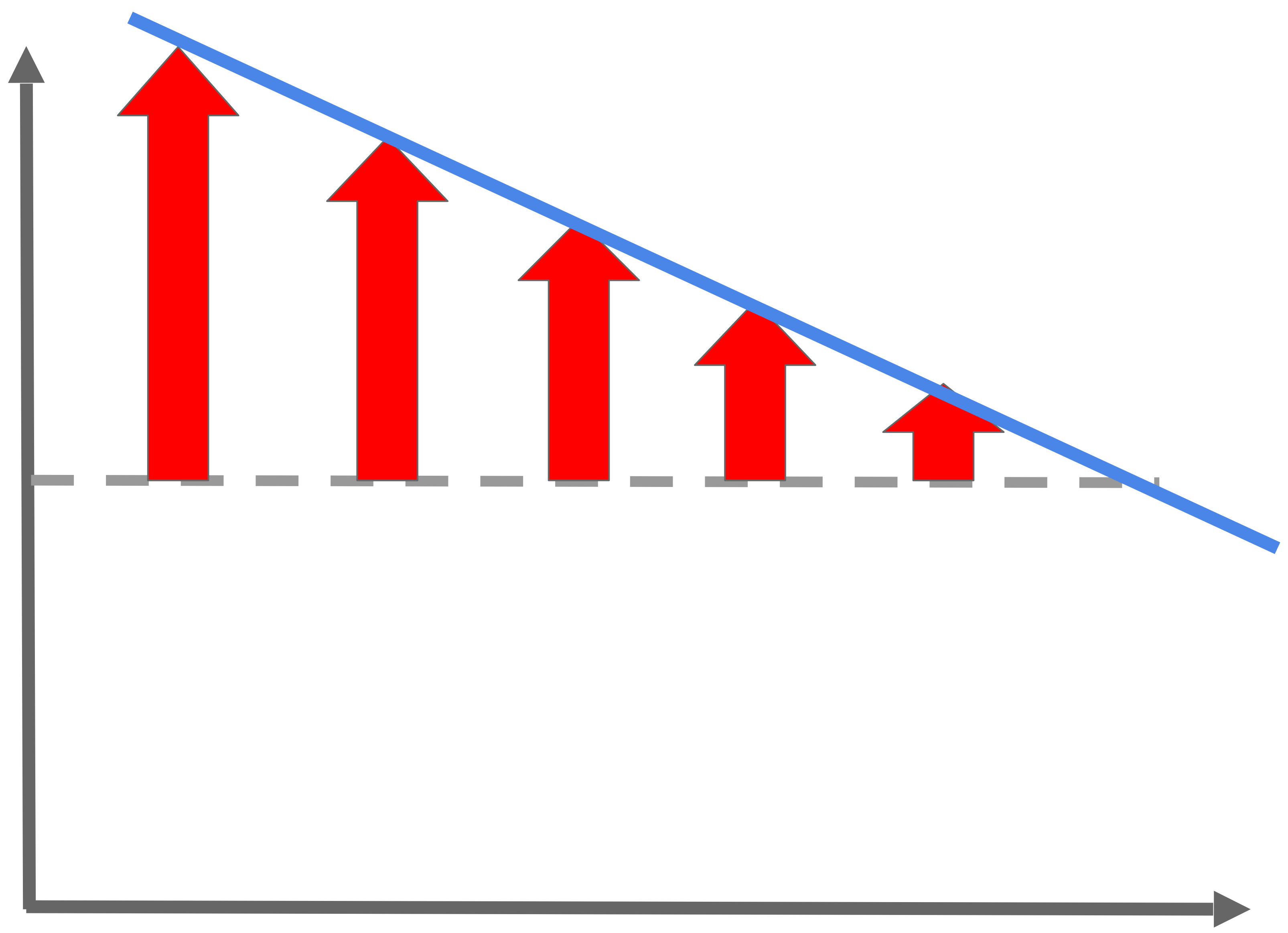
Prime$_{DO}$ → Target$_{PD}$



Increasing PD Biases

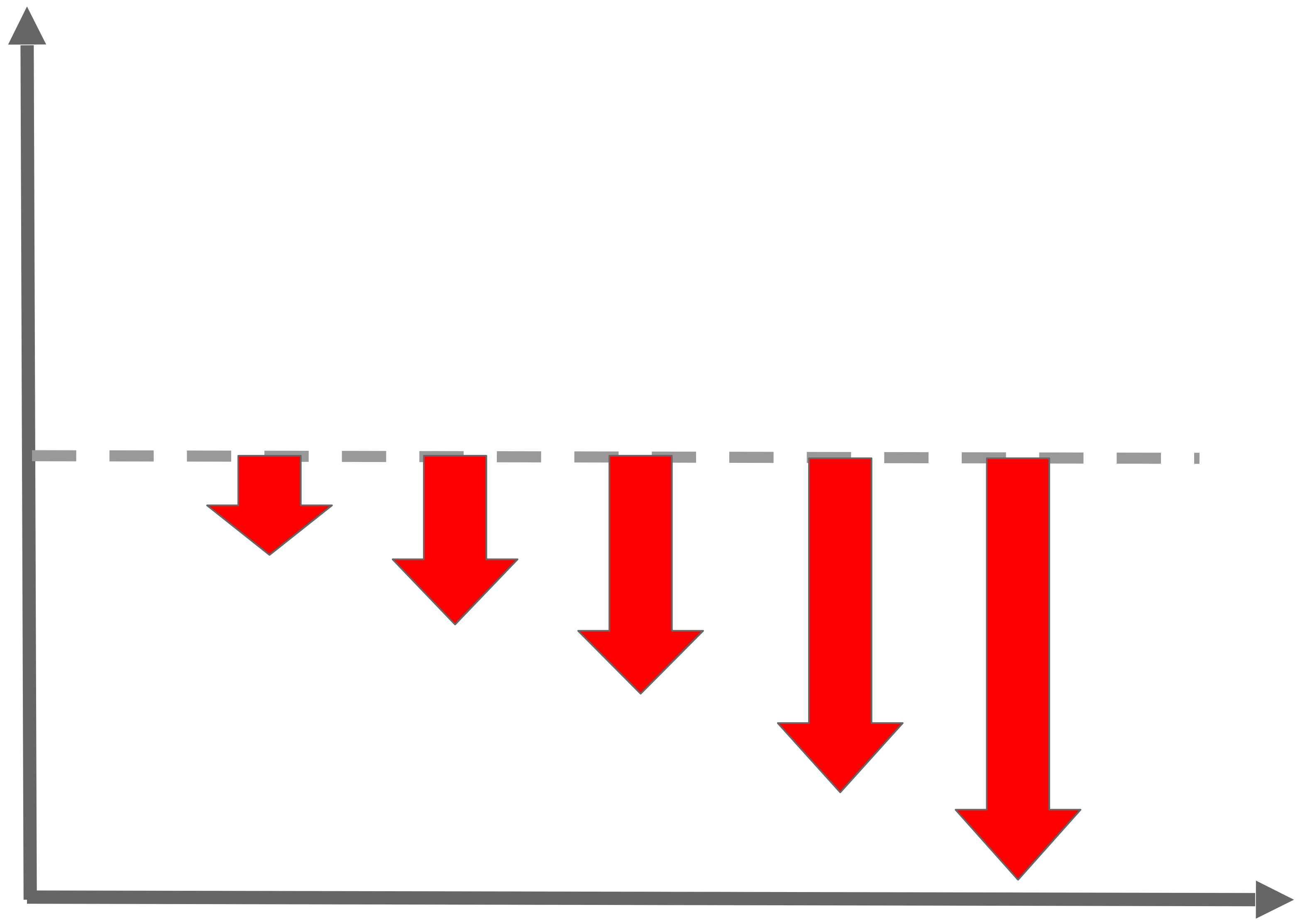
Predictions on the IFE

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



Increasing PD Biases

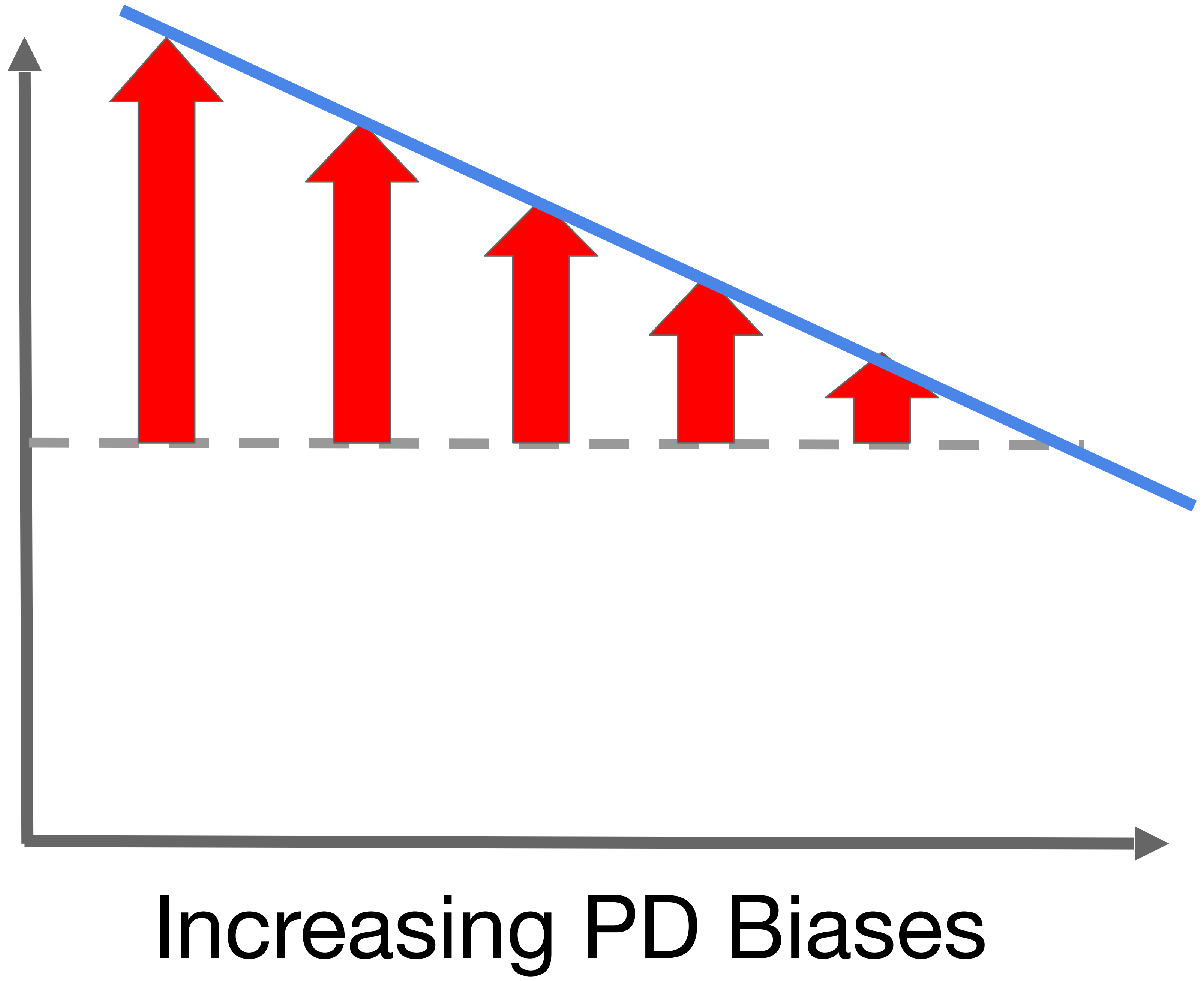
$\text{Prime}_{DO} \rightarrow \text{Target}_{PD}$



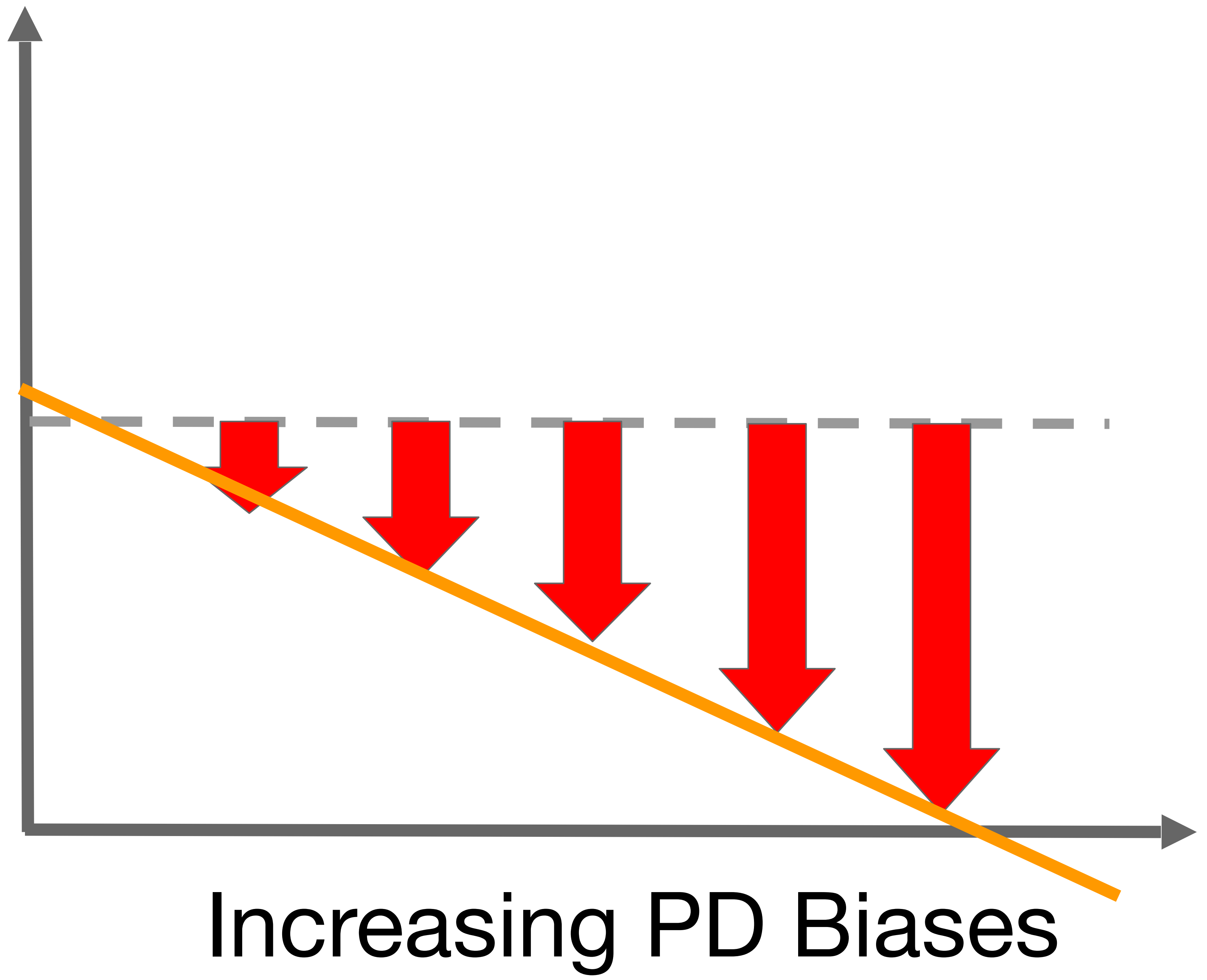
Increasing PD Biases

Predictions on the IFE

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

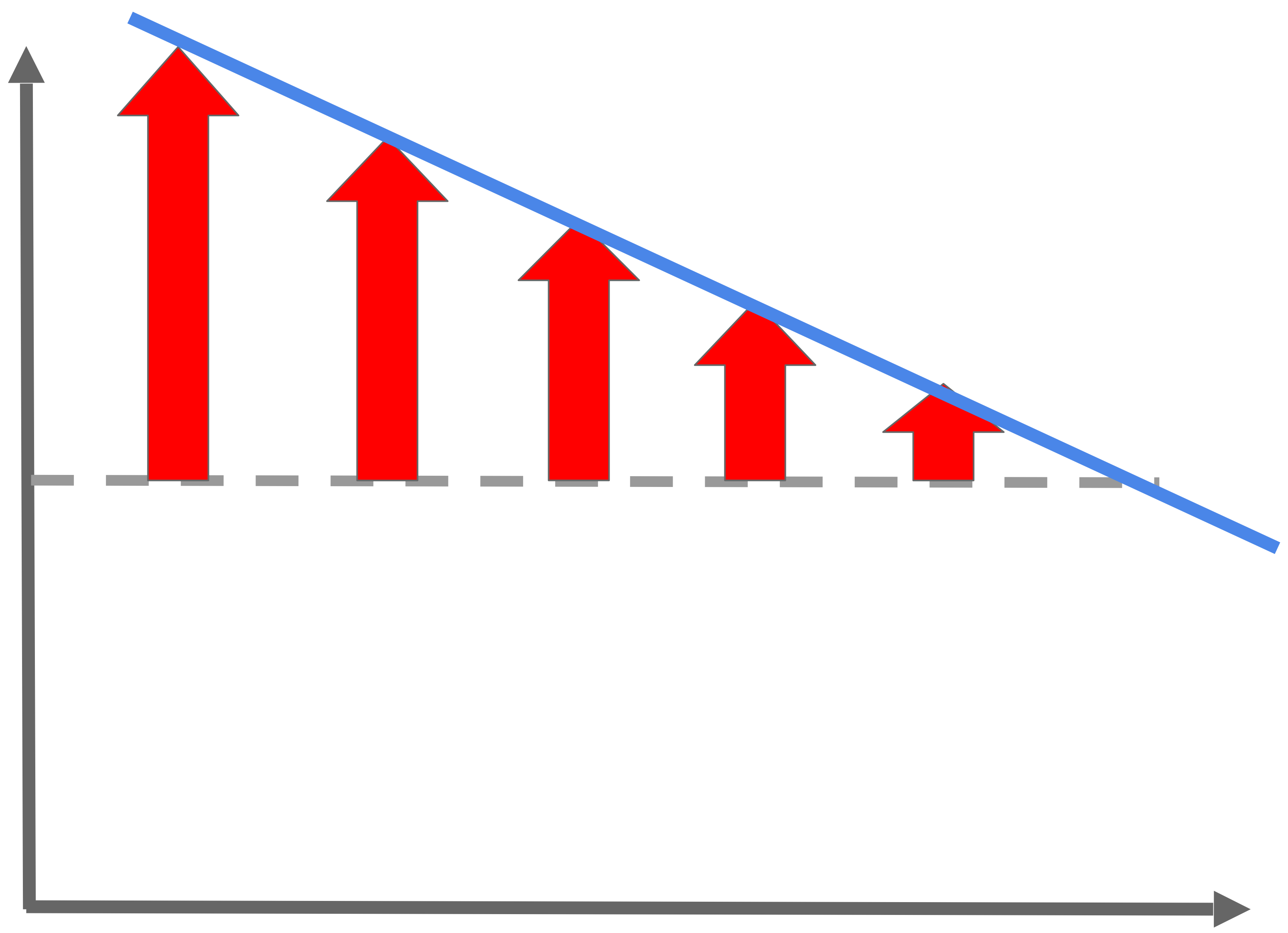


$\text{Prime}_{DO} \rightarrow \text{Target}_{PD}$



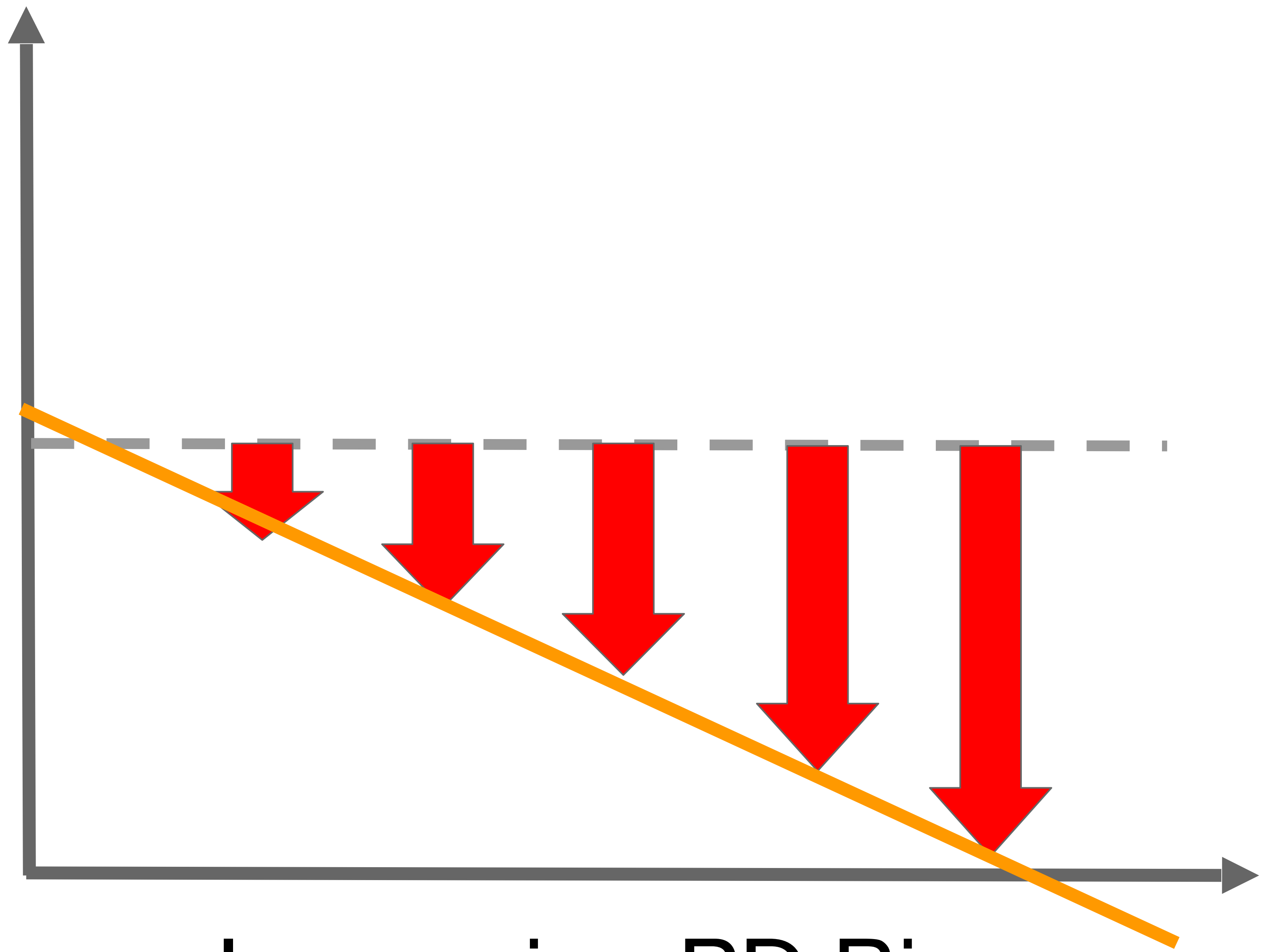
Predictions on the IFE

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



Increasing PD Biases

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



Increasing PD Biases

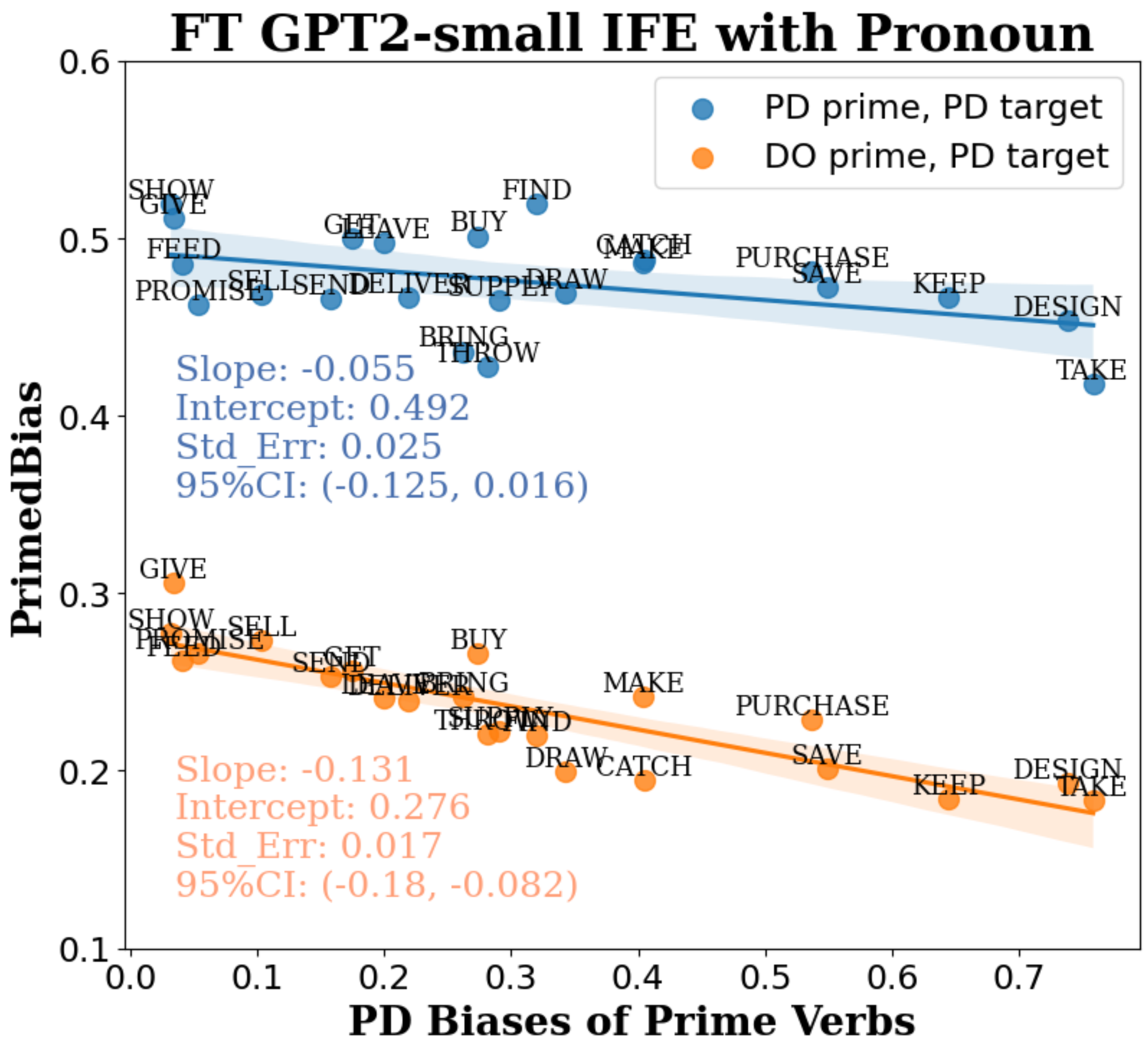
- **IFE**: double negative slopes;
- **Standard Priming**: PD-PD has higher intercept than DO-PD;

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;

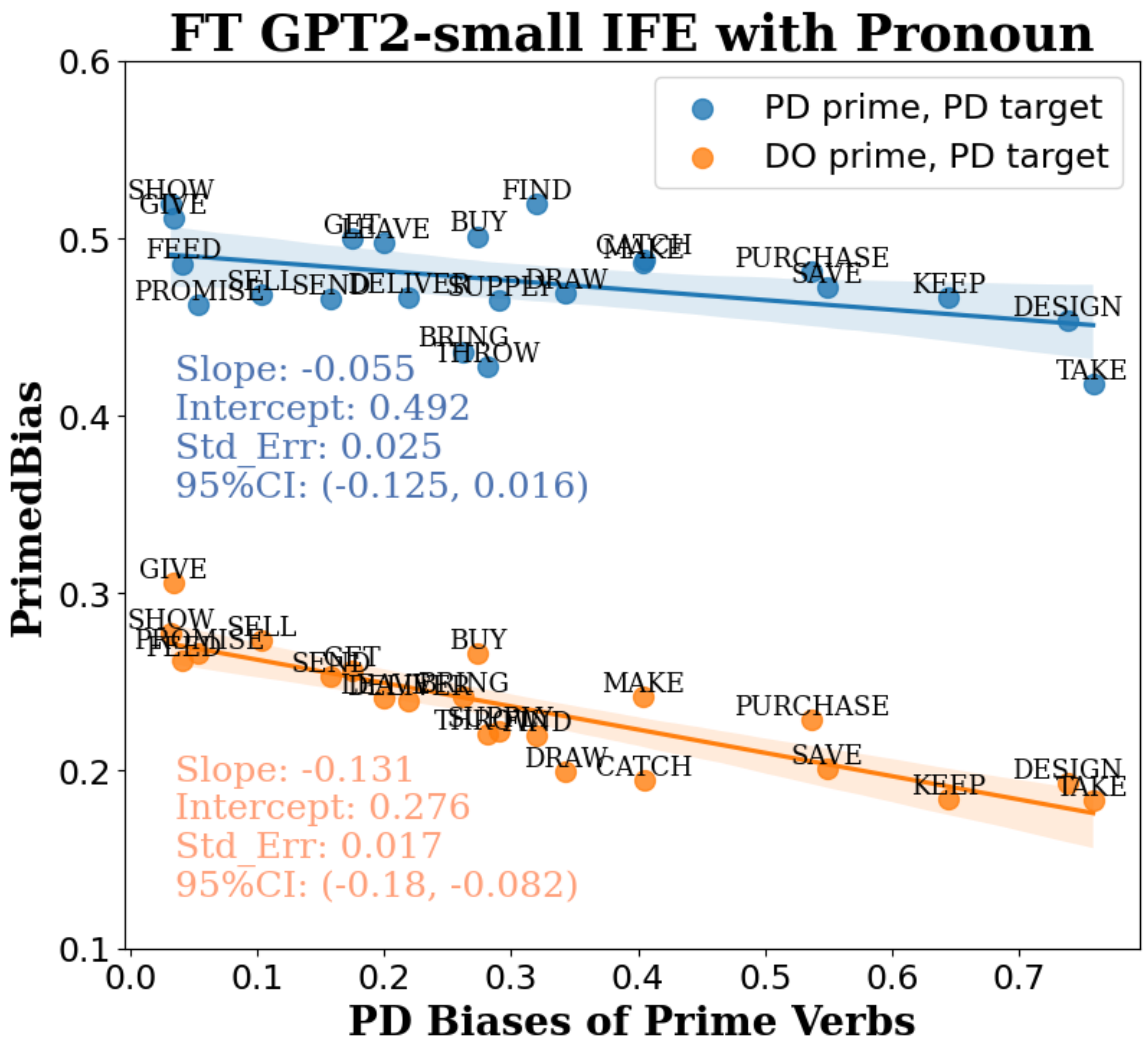
Priming in LLMs

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;



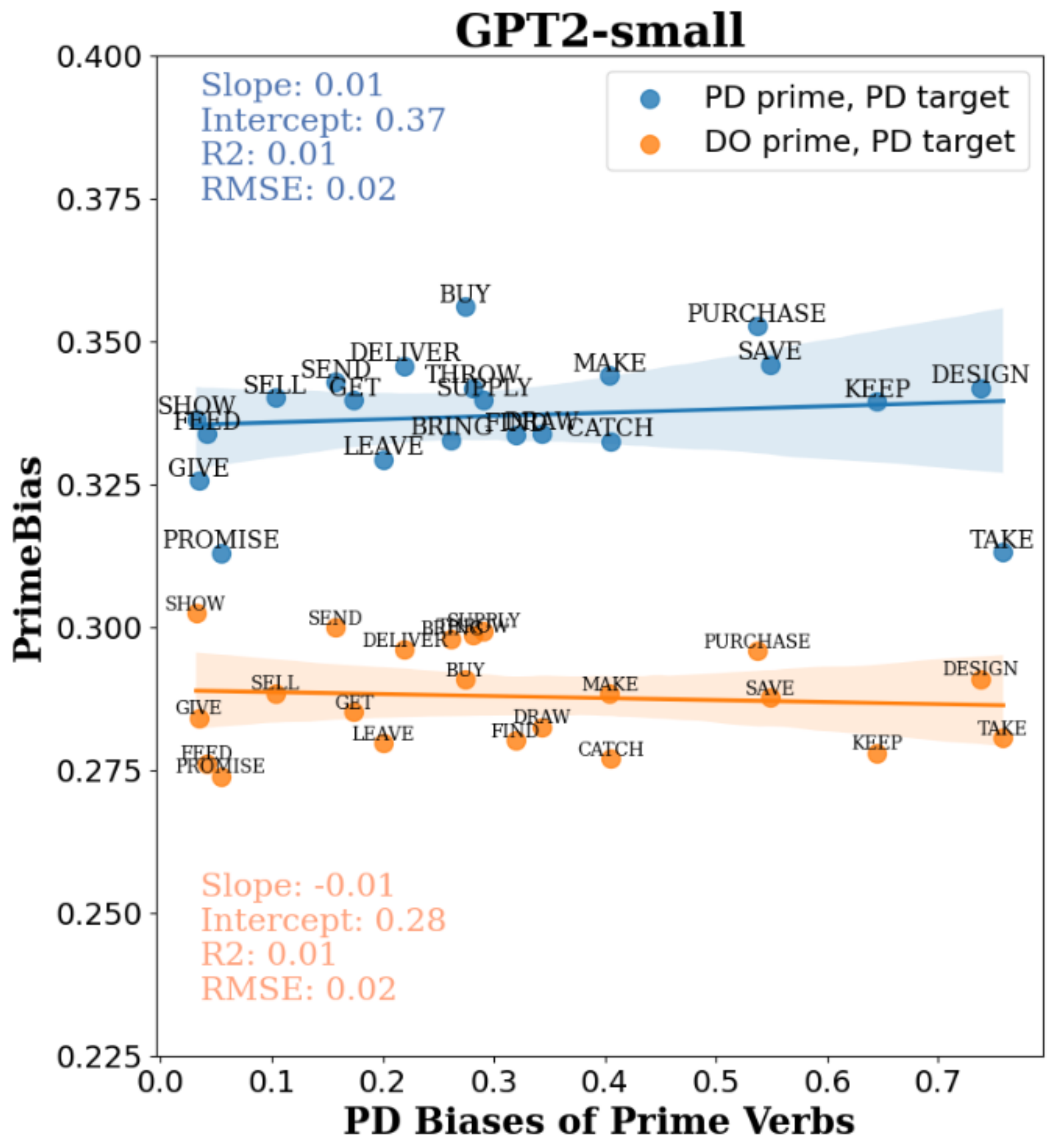
Even *GPT2-small* shows significant inverse frequency effects!

Priming in LLMs

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

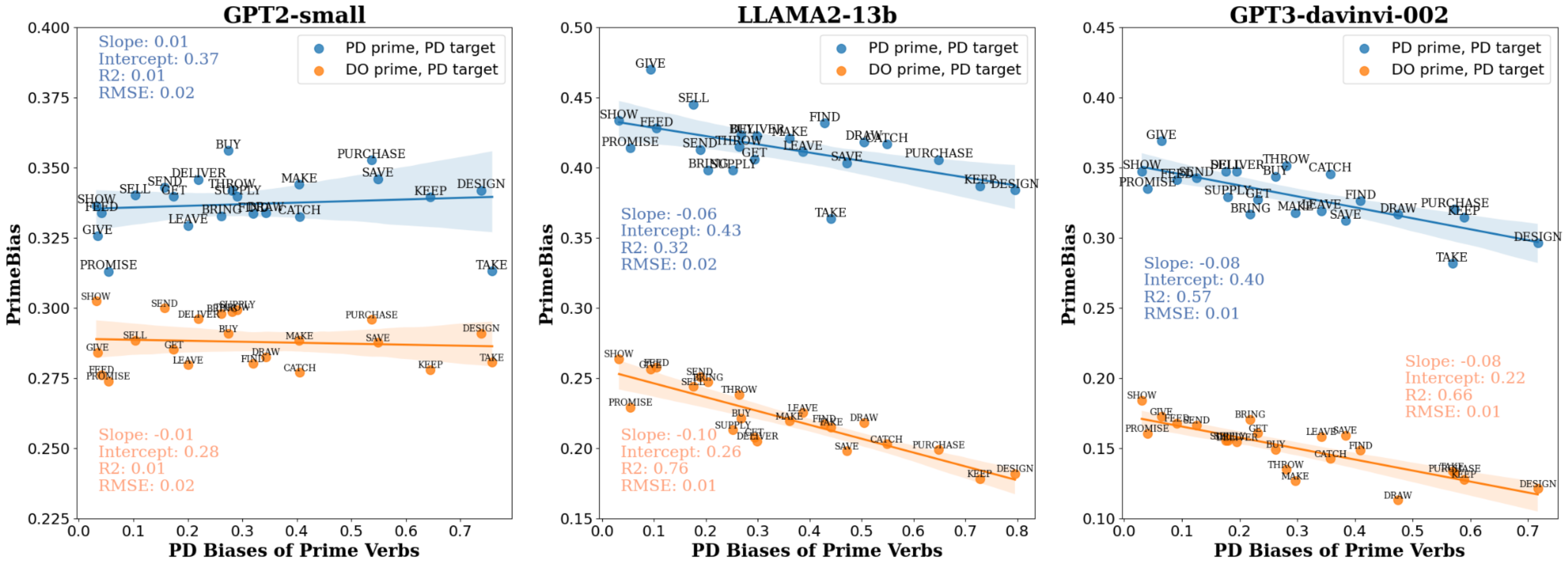
Priming in LLMs

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



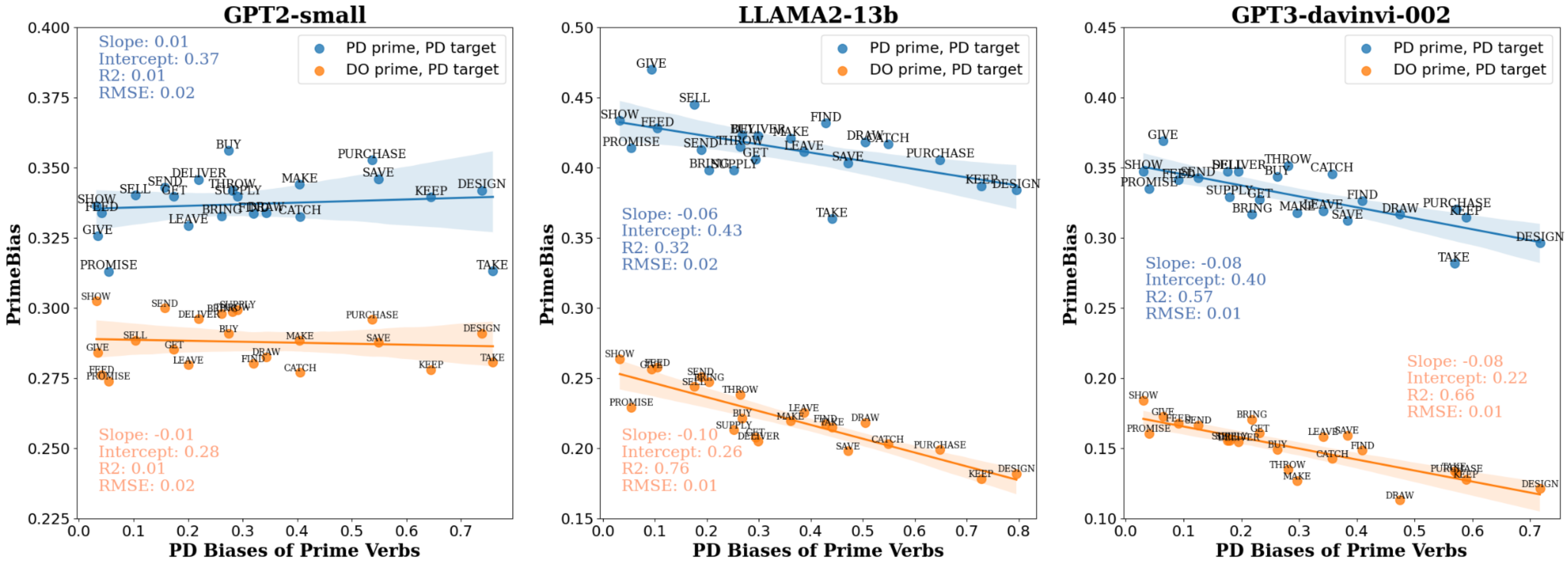
Priming in LLMs

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



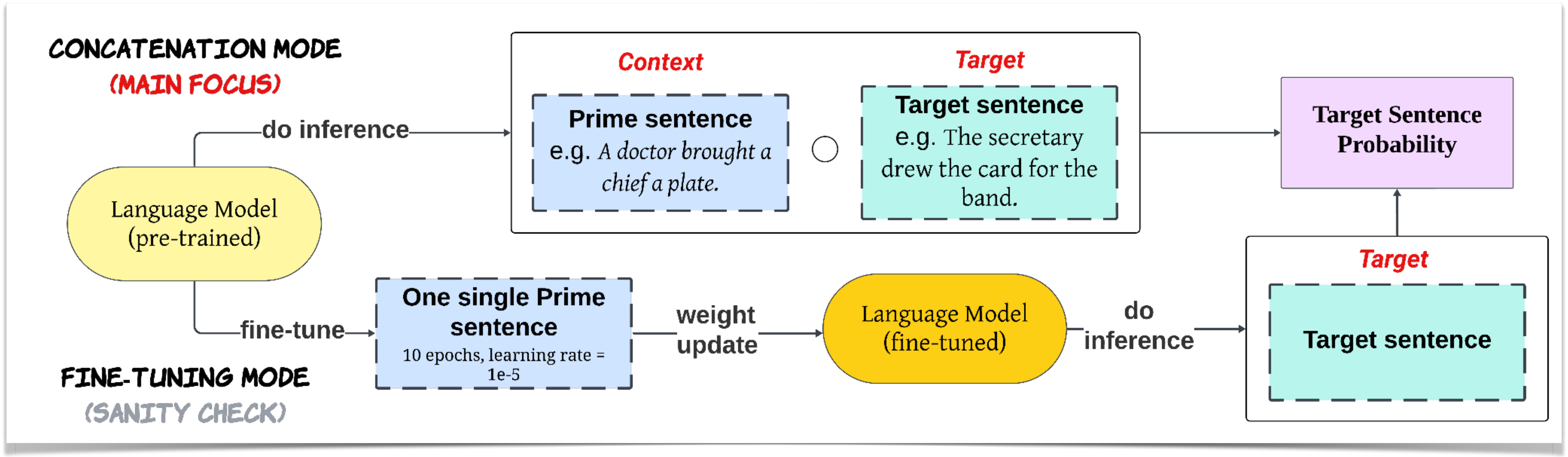
Priming in LLMs

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

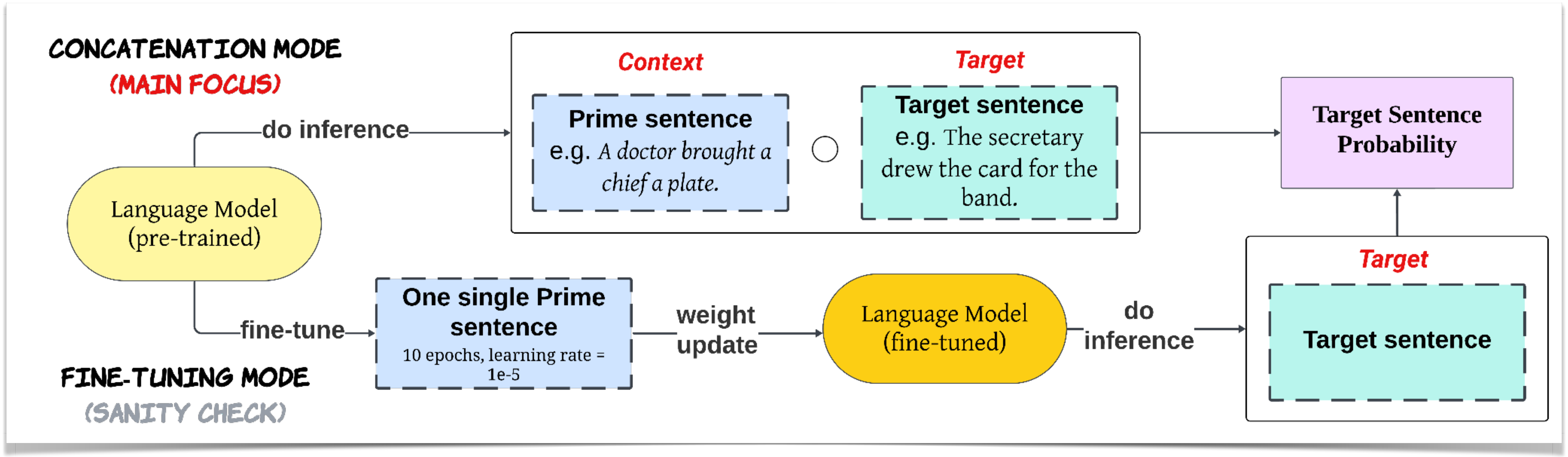


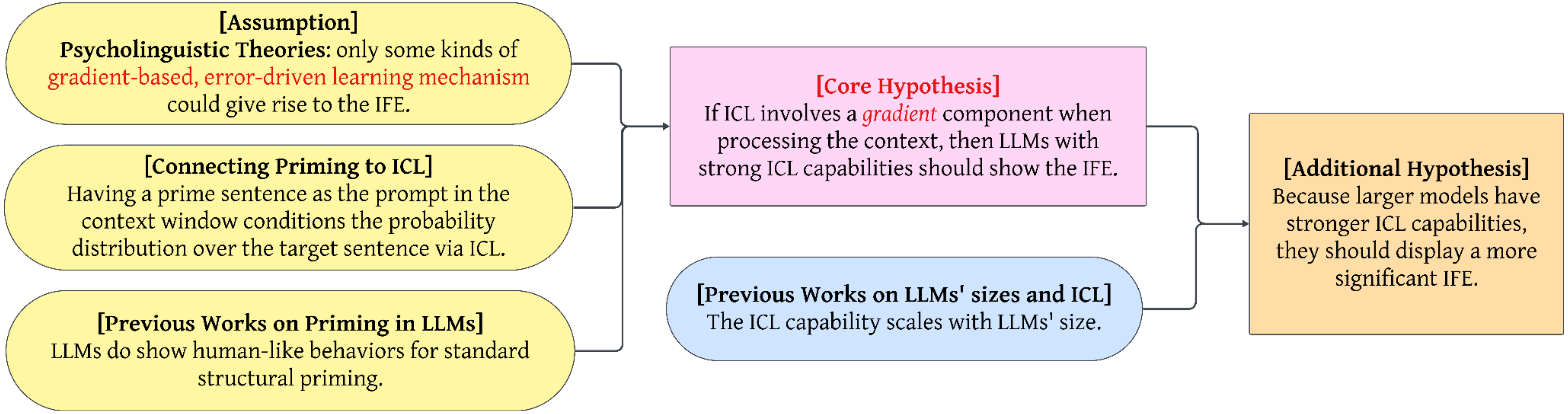
Larger models show more significant inverse frequency effects!

Recap

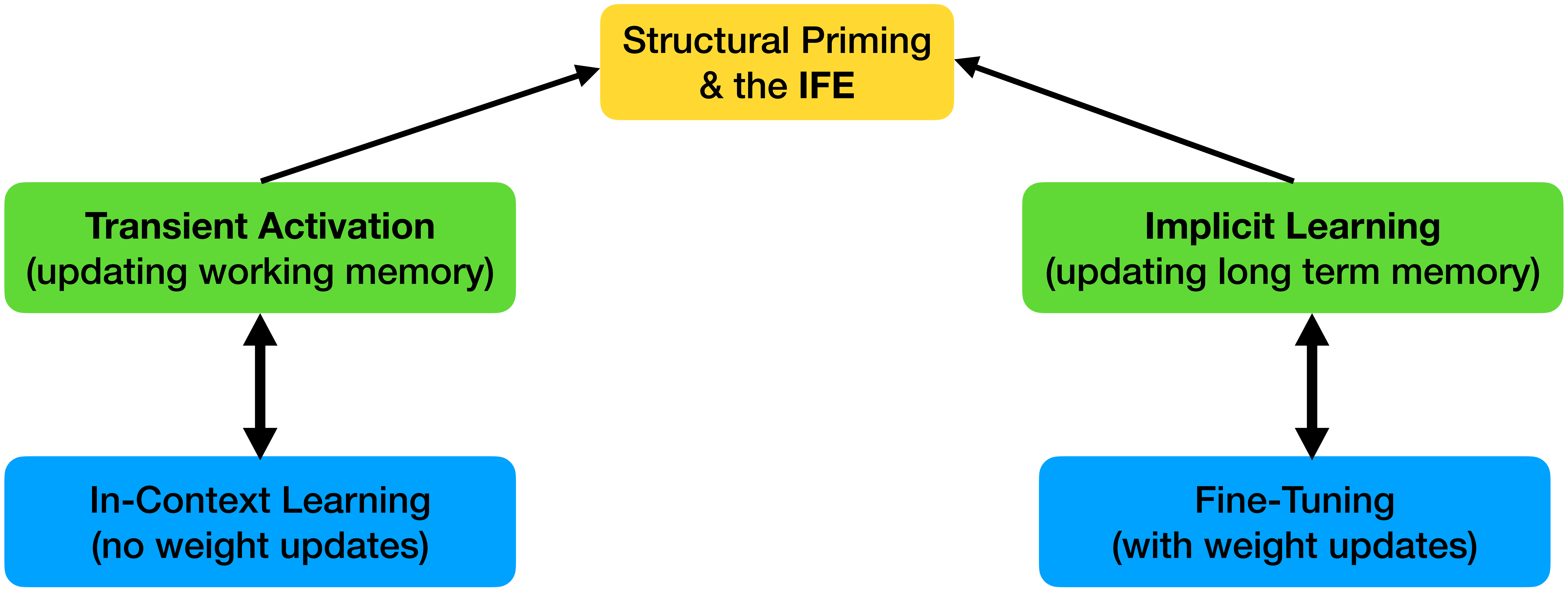

CONCATENATION MODE
(MAIN FOCUS)
do inference
Context
Target
Prime sentence
e.g. A doctor brought a chief a plate.
Target sentence
e.g. The secretary drew the card for the band.
Target Sentence Probability
Language Model (pre-trained)
fine-tune
One single Prime sentence
10 epochs, learning rate = 1e-5
weight update
Language Model (fine-tuned)
do inference
Target
Target sentence
FINE-TUNING MODE
(SANITY CHECK)

Recap

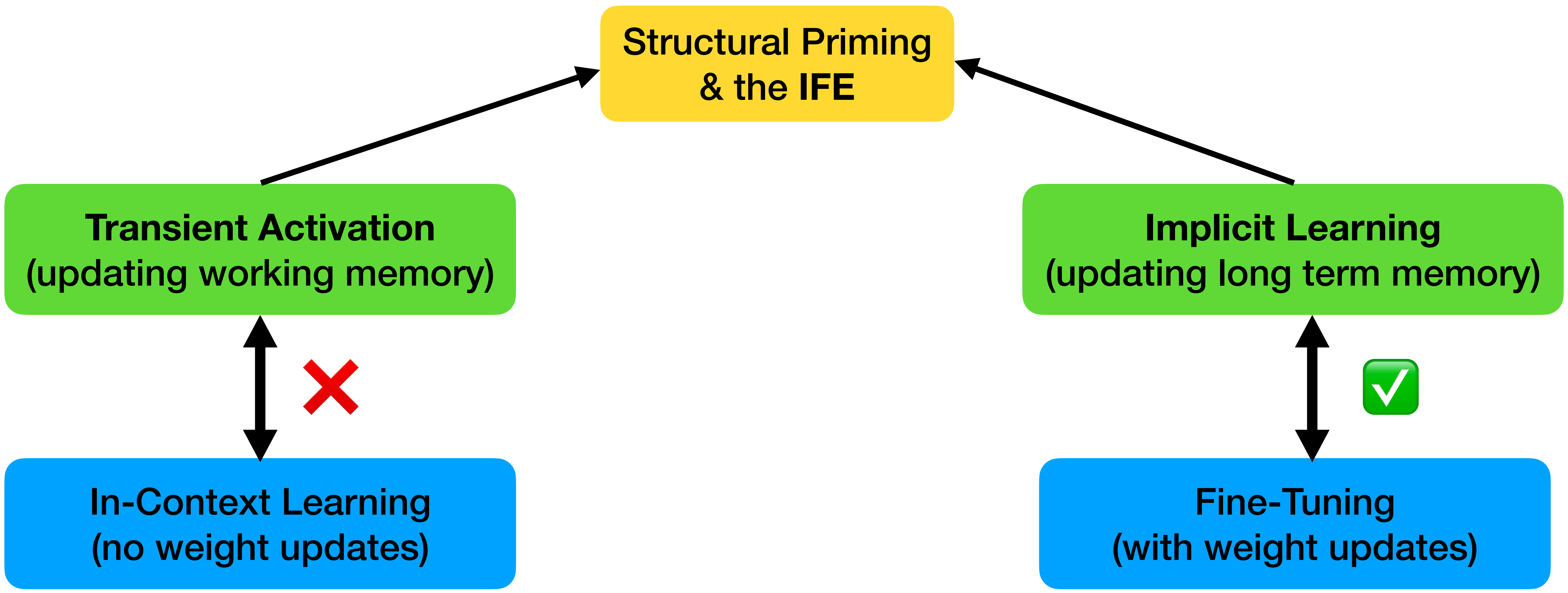

CONCATENATION MODE
(MAIN FOCUS)
do inference
Language Model
(pre-trained)
Context
Prime sentence
e.g. A doctor brought a chief a plate.
Target
Target sentence
e.g. The secretary drew the card for the band.
Target Sentence Probability
fine-tune
One single Prime sentence
10 epochs, learning rate = 1e-5
weight update
Language Model
(fine-tuned)
do inference
Target
Target sentence
FINE-TUNING MODE
(SANITY CHECK)


[Assumption]
Psycholinguistic Theories: only some kinds of gradient-based, error-driven learning mechanism could give rise to the IFE.
[Connecting Priming to ICL]
Having a prime sentence as the prompt in the context window conditions the probability distribution over the target sentence via ICL.
[Previous Works on Priming in LLMs]
LLMs do show human-like behaviors for standard structural priming.
[Core Hypothesis]
If ICL involves a gradient component when processing the context, then LLMs with strong ICL capabilities should show the IFE.
[Previous Works on LLMs' sizes and ICL]
The ICL capability scales with LLMs' size.
[Additional Hypothesis]
Because larger models have stronger ICL capabilities, they should display a more significant IFE.

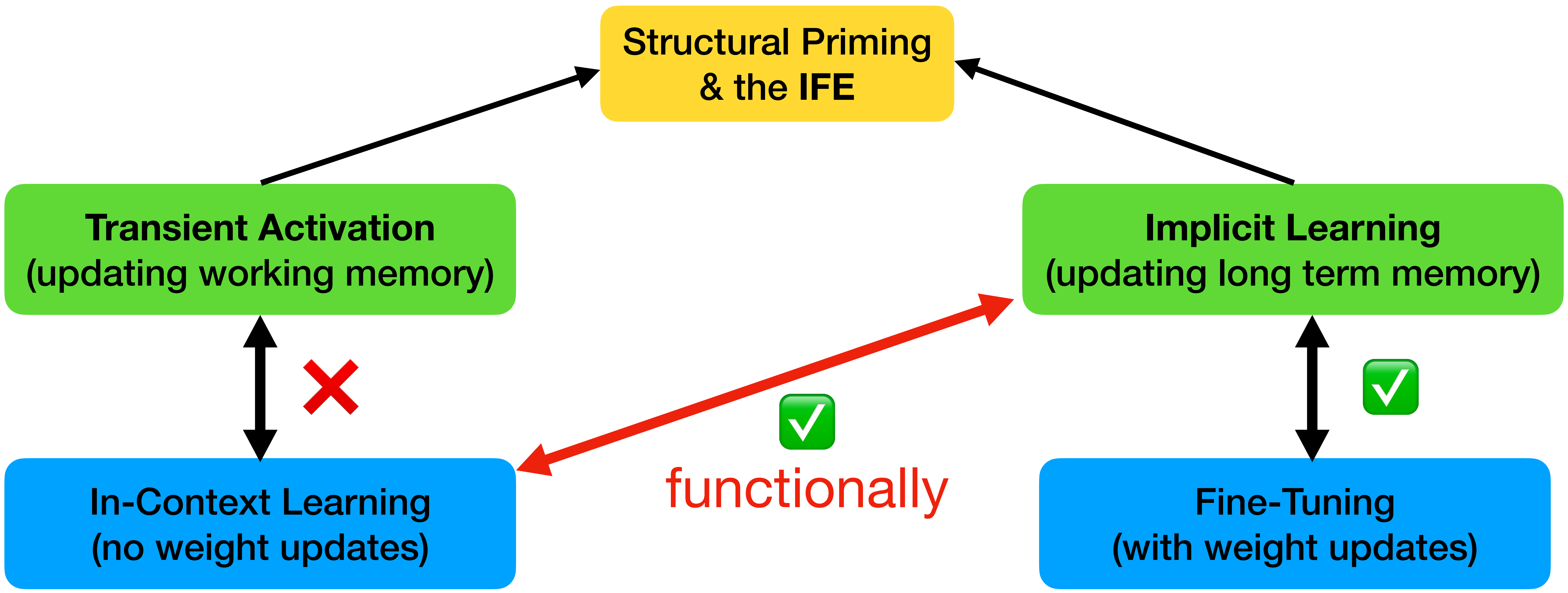
Takeaways from Experiment 1



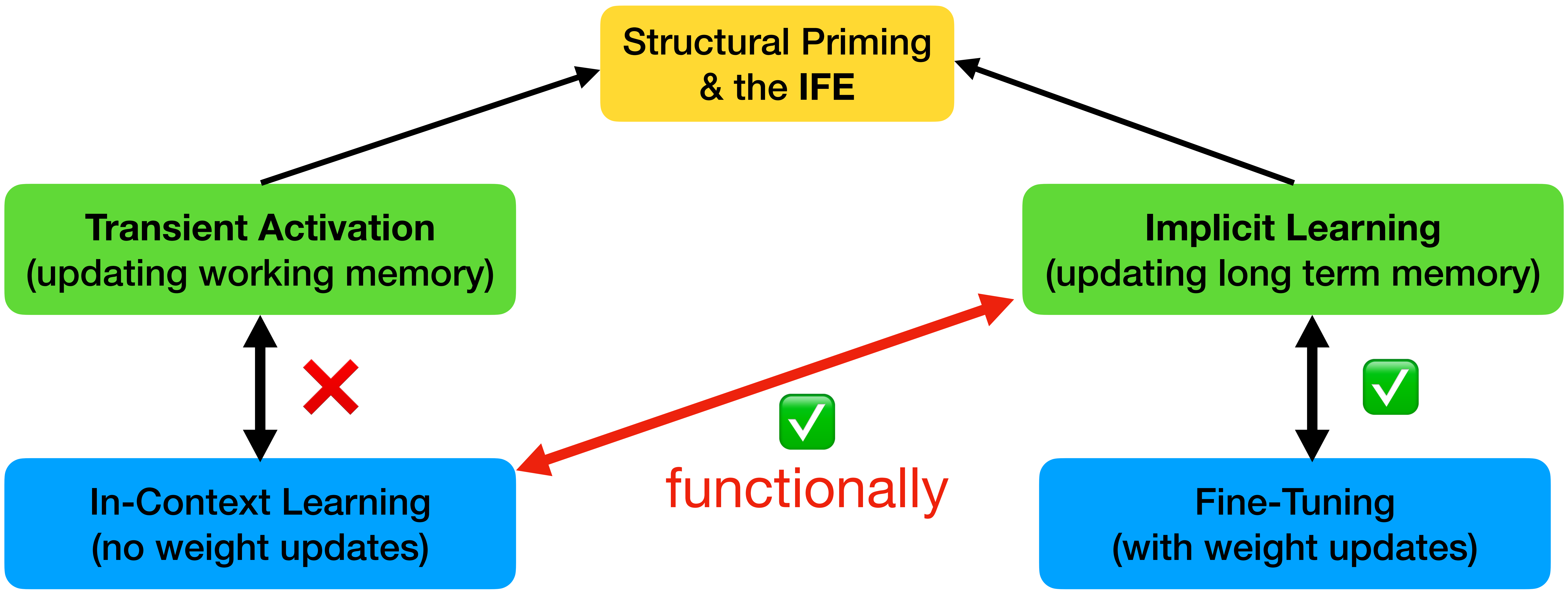
Takeaways from Experiment 1



Takeaways from Experiment 1



Takeaways from Experiment 1



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:

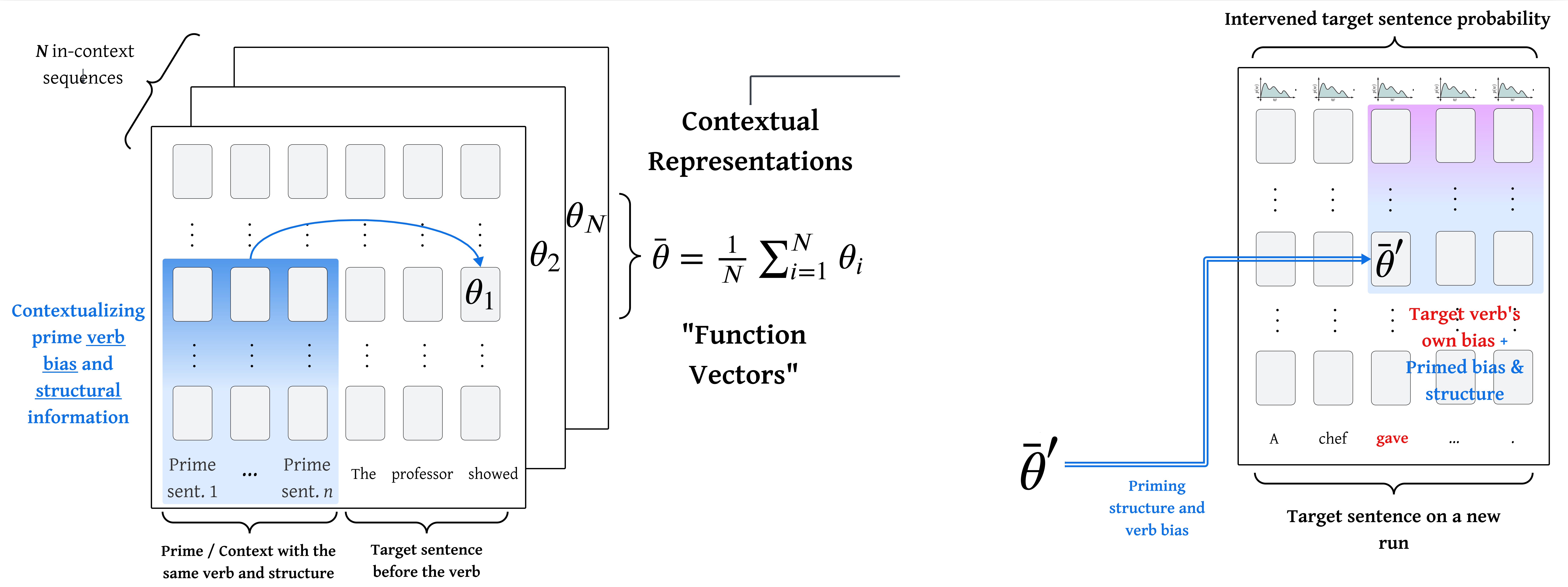
- Then, what is truly equivalent to “transient activation” in LLMs?
- How could we causally verify the involvement of verb bias in IFE?

Research Questions:

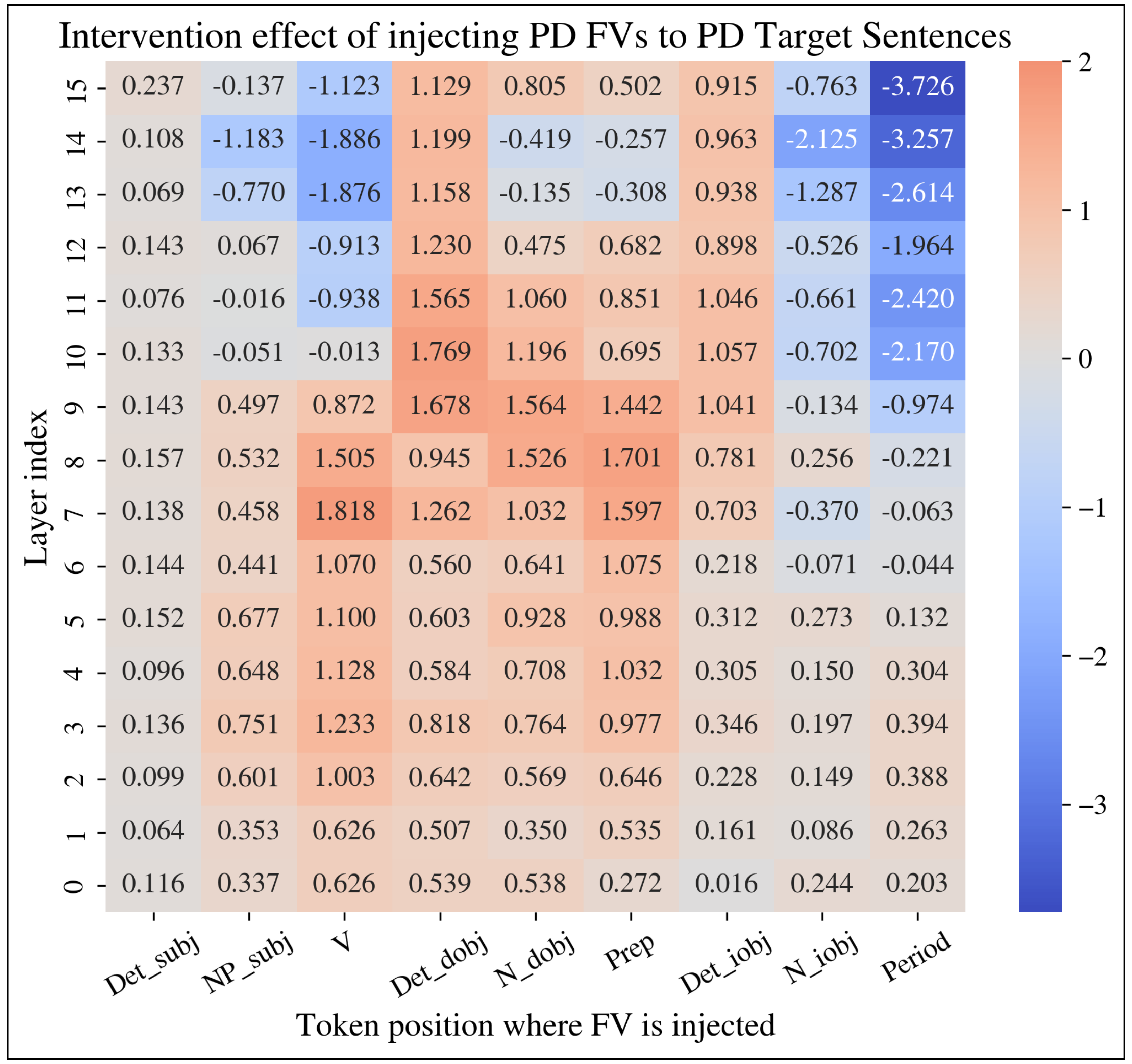
- Then, what is truly equivalent to “transient activation” in LLMs?
- How could we causally verify the involvement of verb bias in IFE?

Let’s go beyond purely behavioral evaluation and open up the black box!

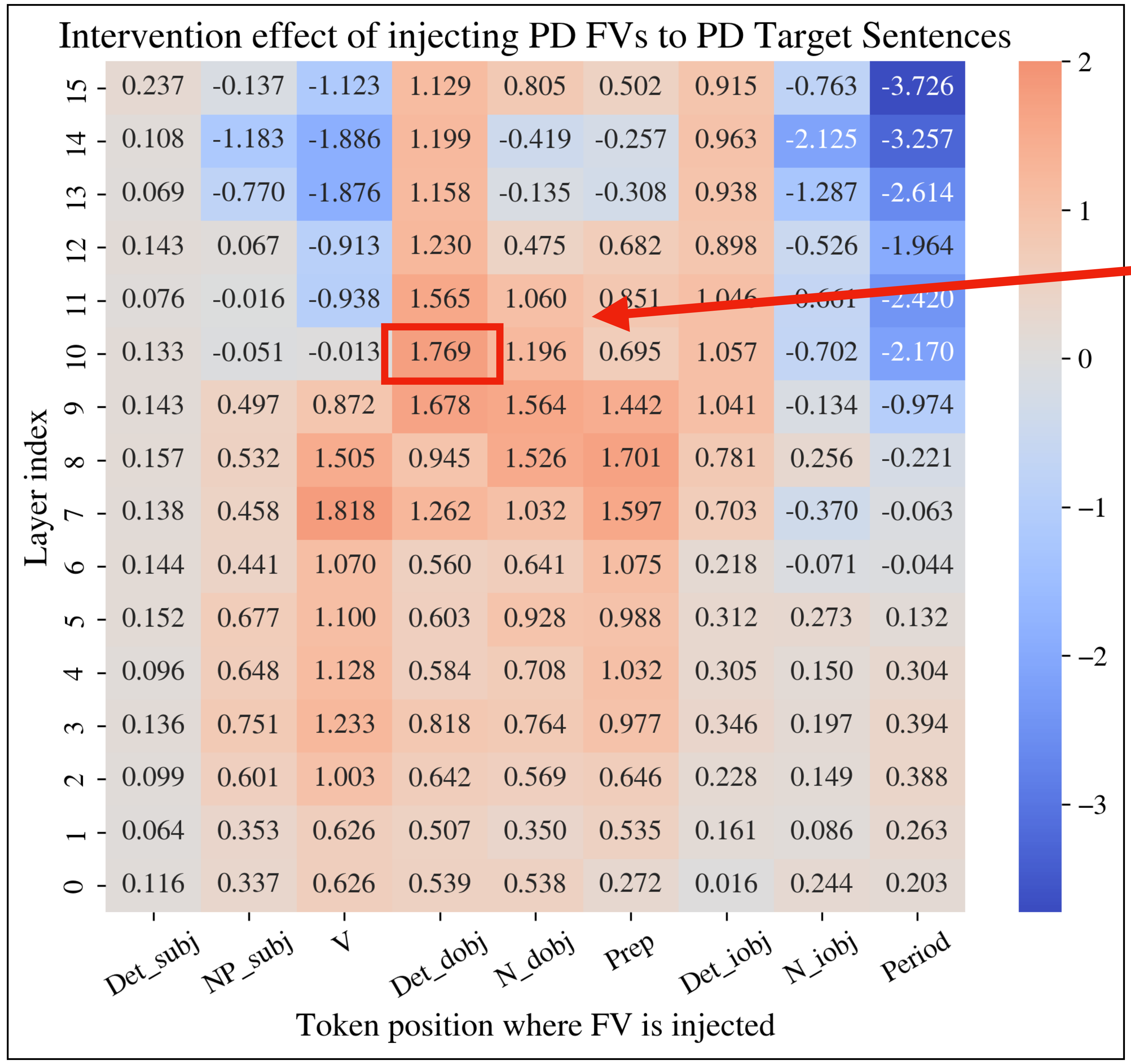
Simulating Transient Activations



Simulating Transient Activations

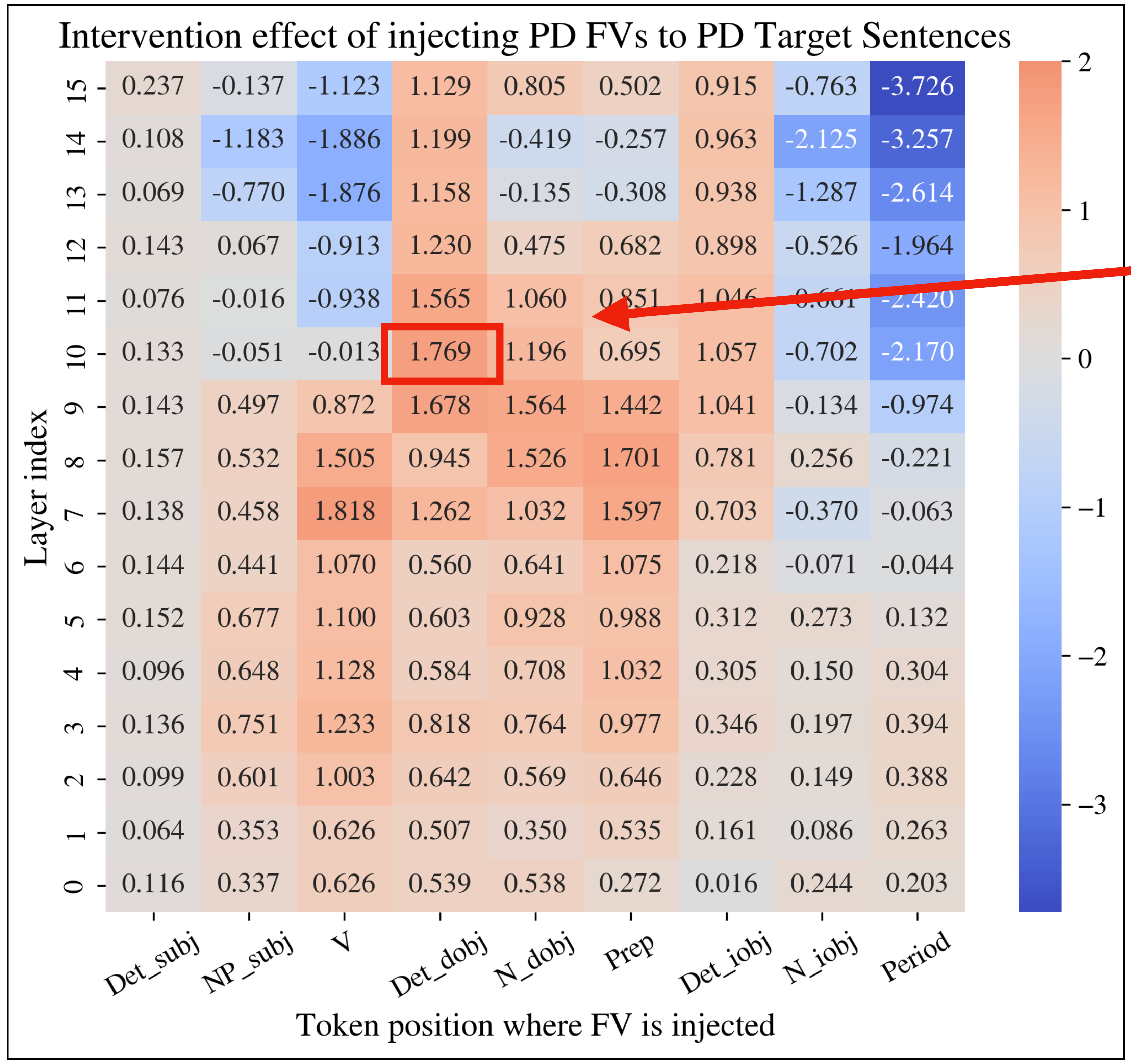


Simulating Transient Activations

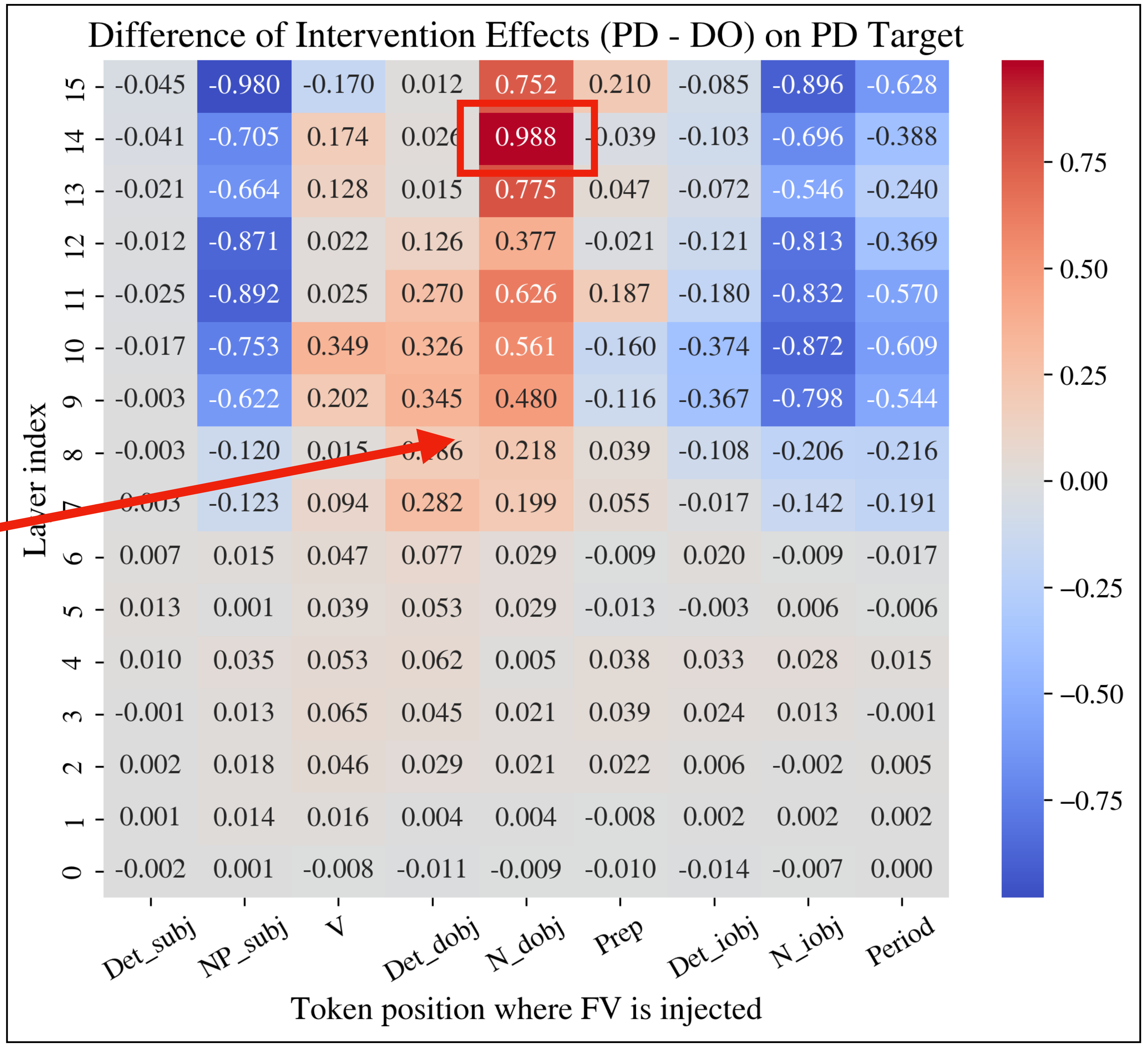


- Observing the priming effect: FVs increased target sentence probability;

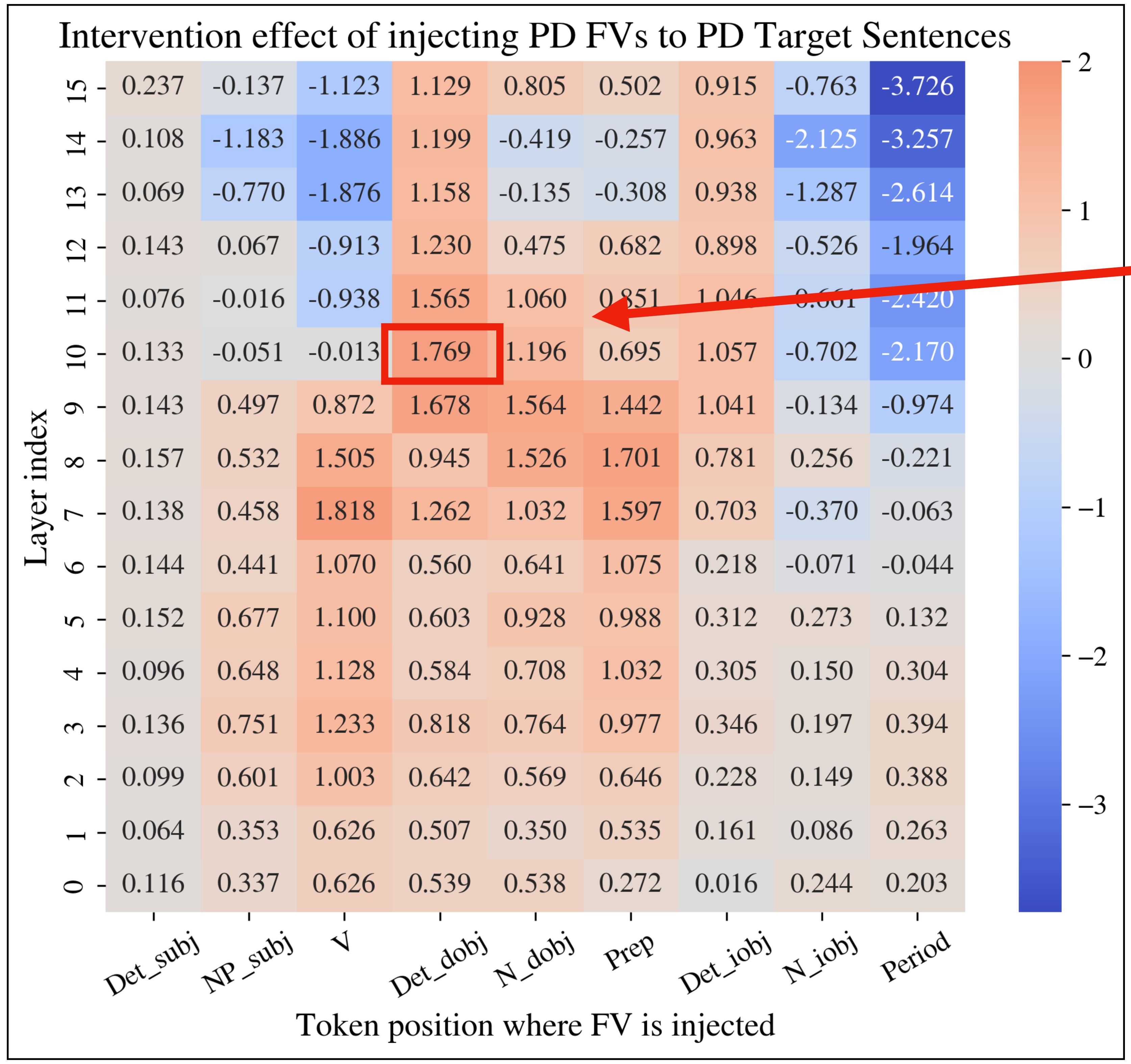
Simulating Transient Activations



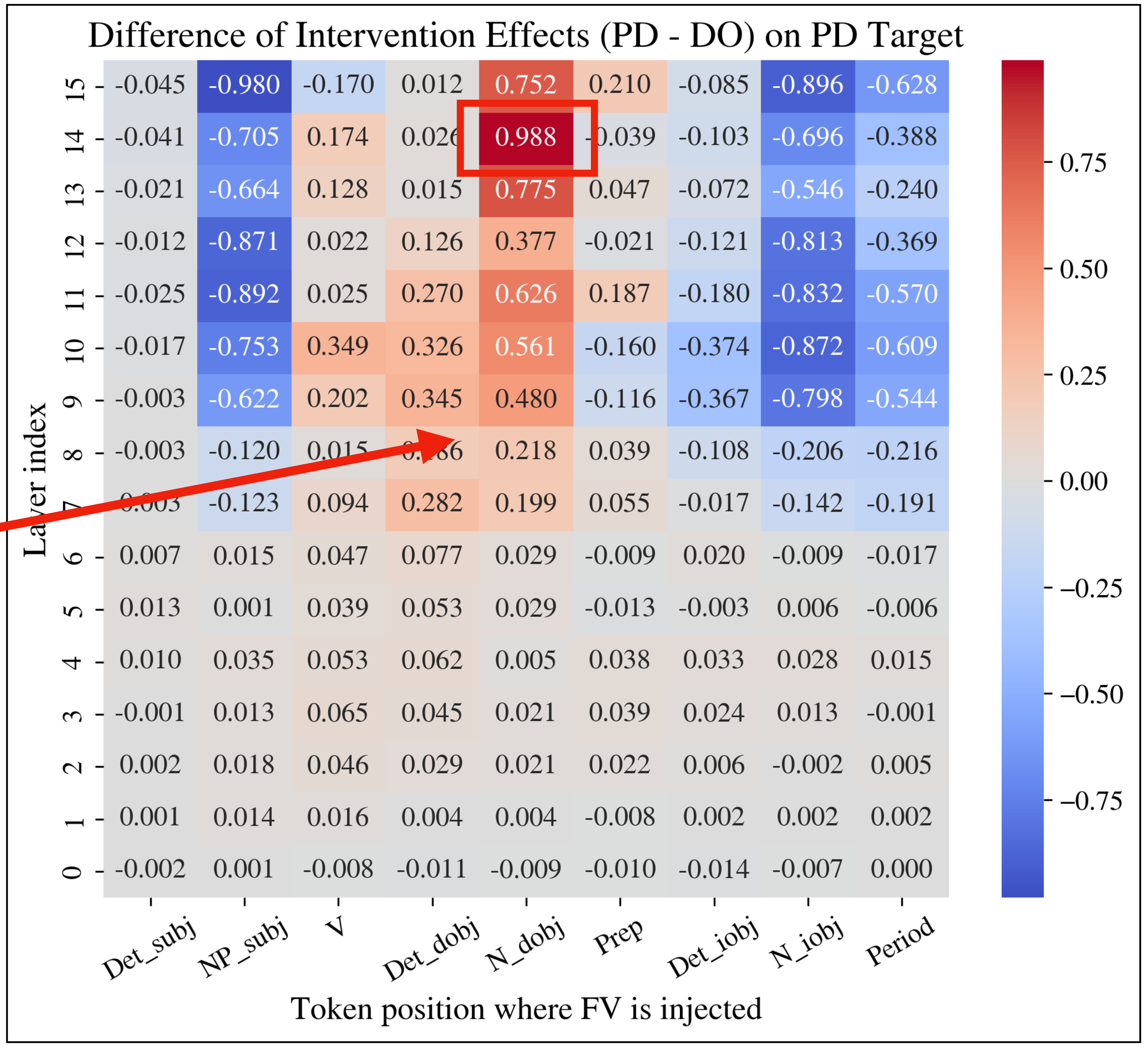
- Observing the priming effect: FVs increased target sentence probability;
- PD FVs cause a larger intervention effect than DO FVs on PD targets.



Simulating Transient Activations

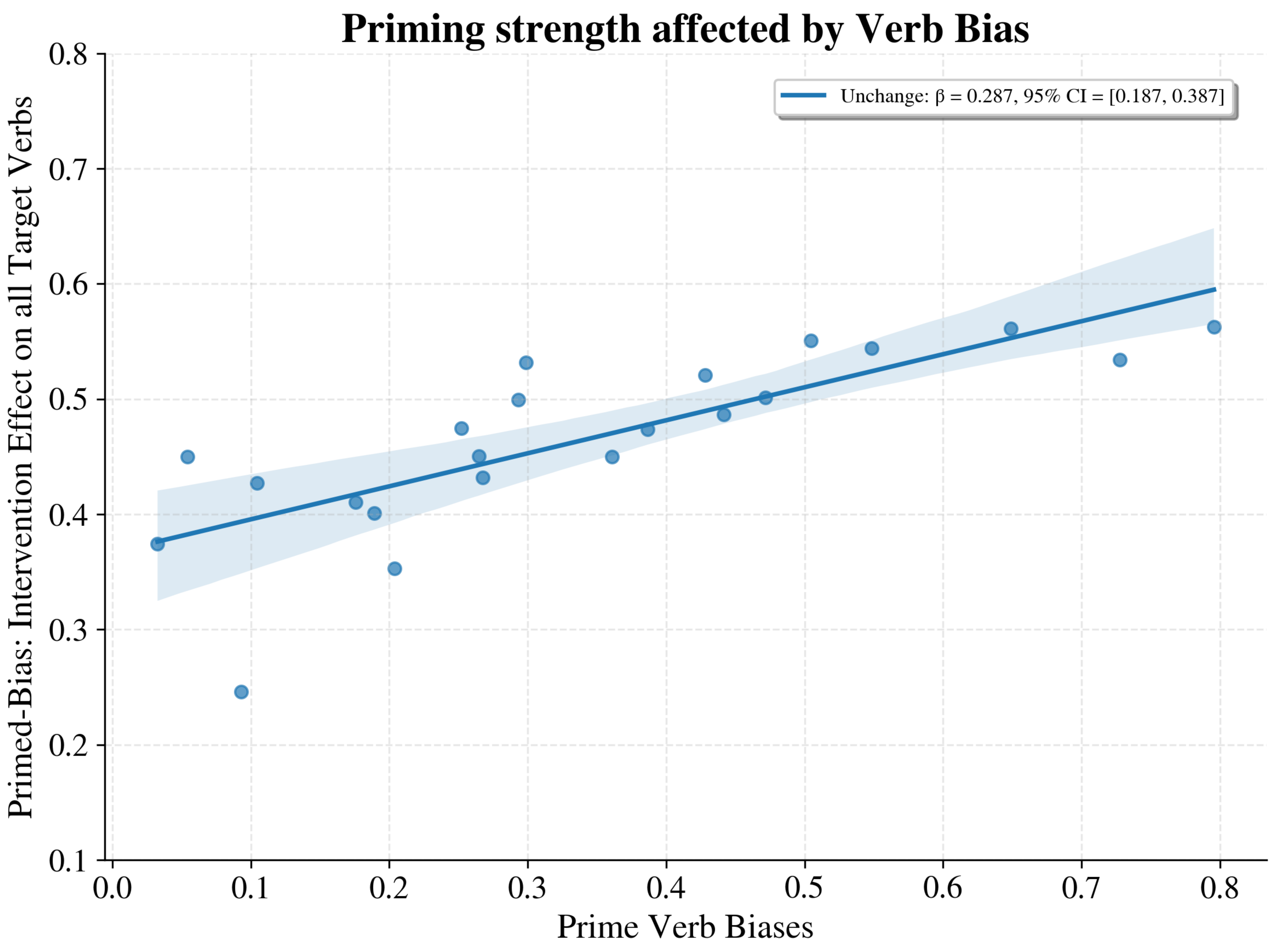


- Observing the priming effect: FVs increased target sentence probability;
- PD FVs cause a larger intervention effect than DO FVs on PD targets.



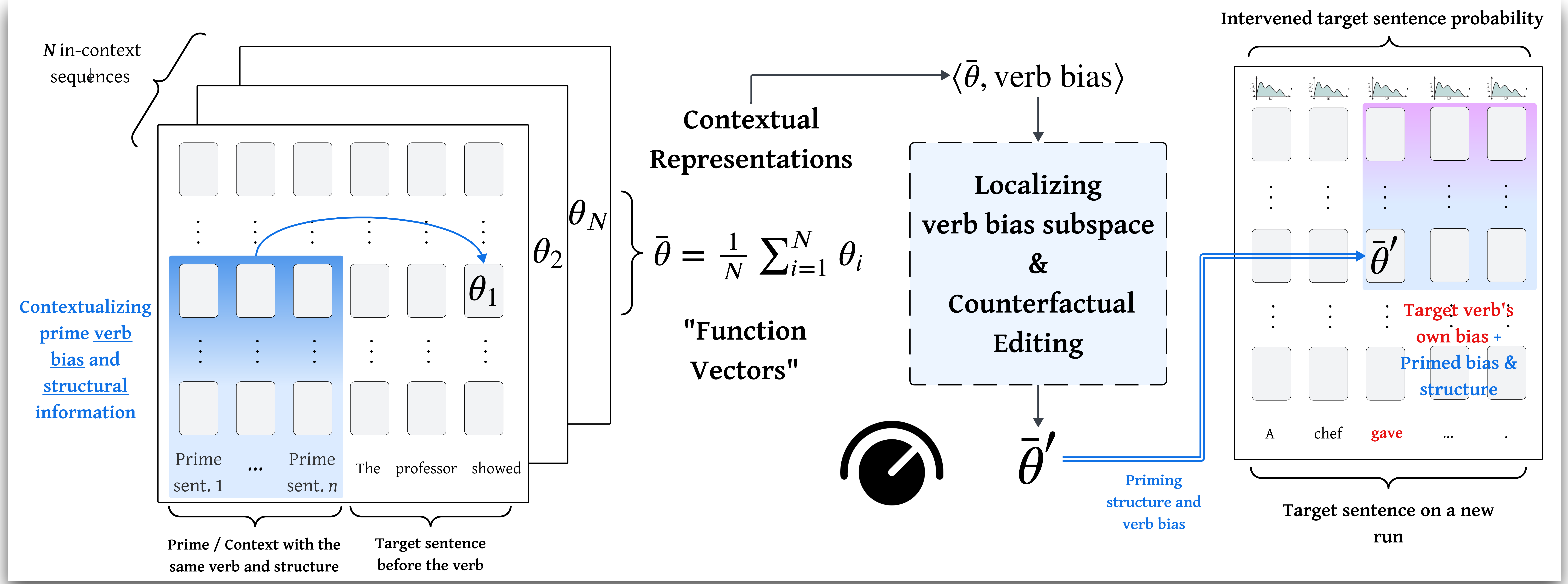
Yes, we do observe standard priming effect.

Simulating Transient Activations

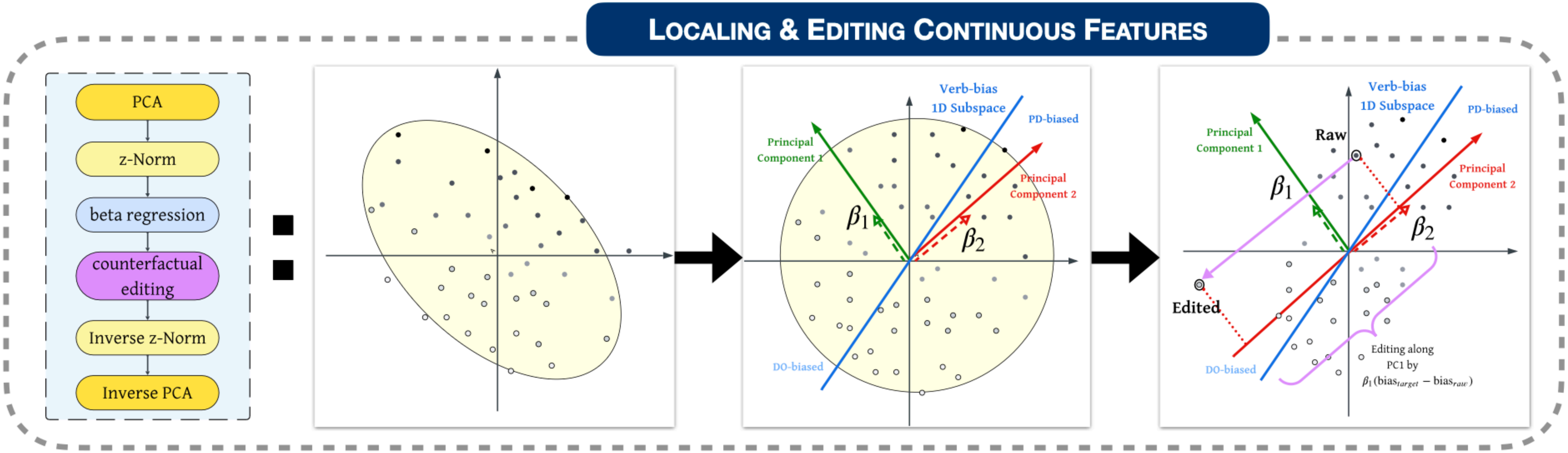


However, no IFE is observed...!

Causal Verification

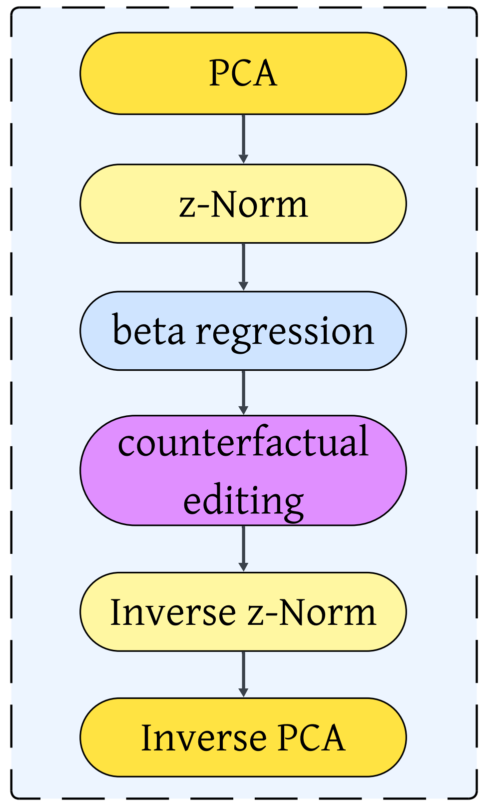


Causal Verification

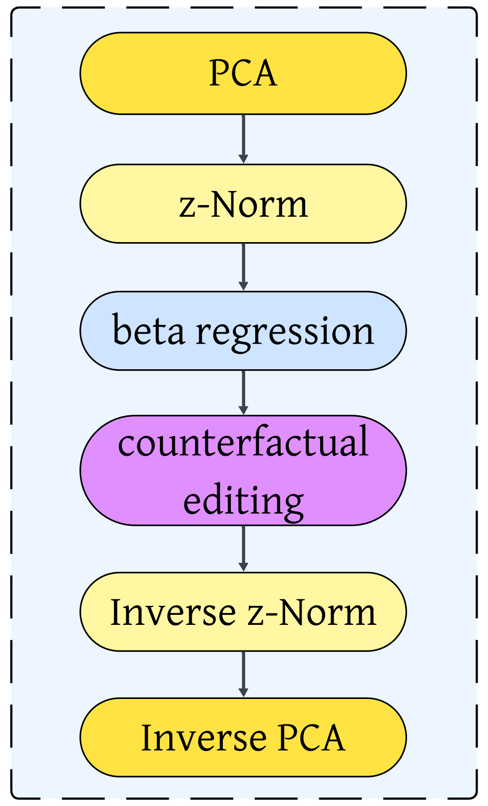


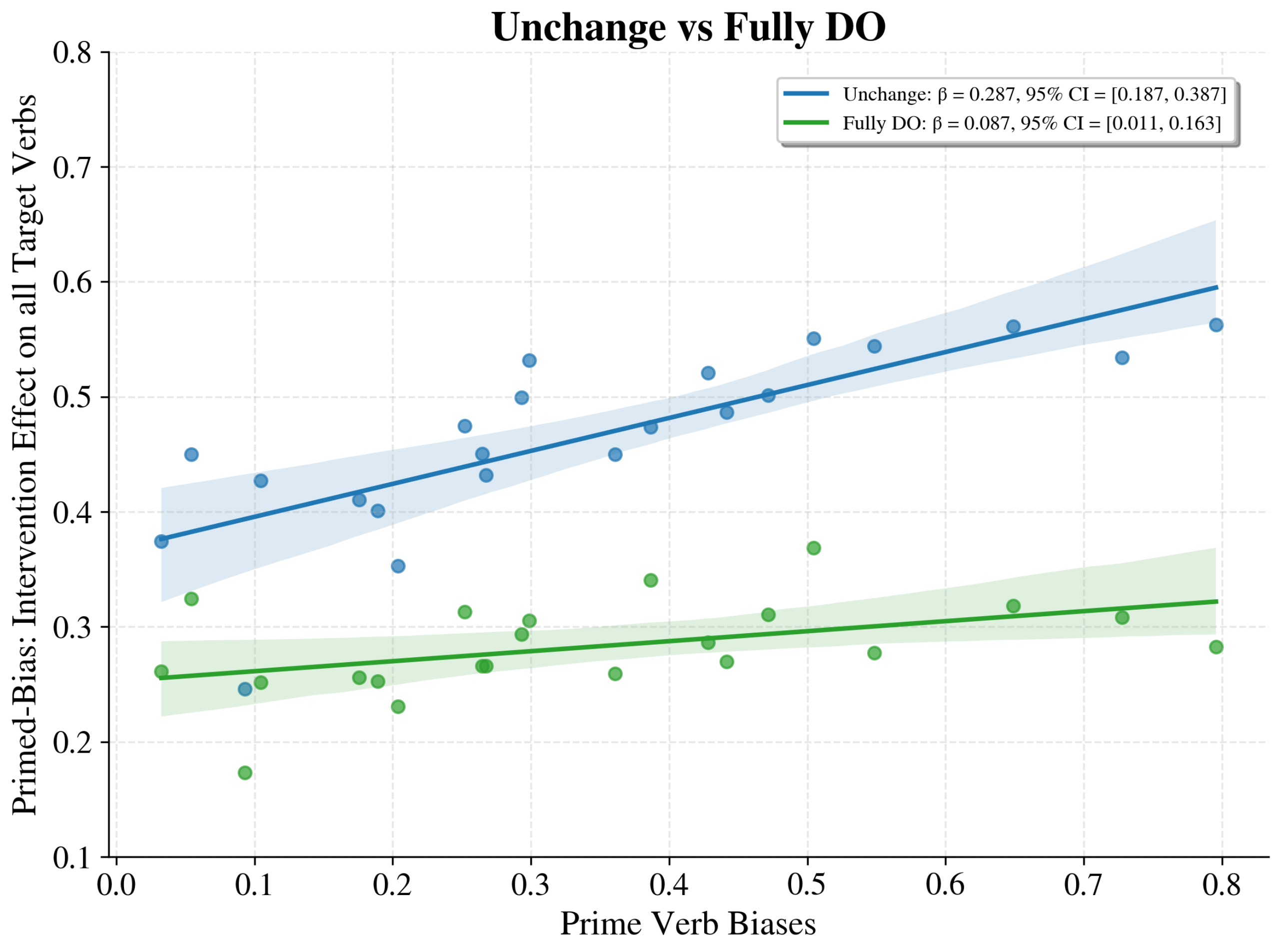
Localizing a 1-dimensional space that represents verb biases, and causally manipulate the verb bias values without changing other information.

Causal Verification

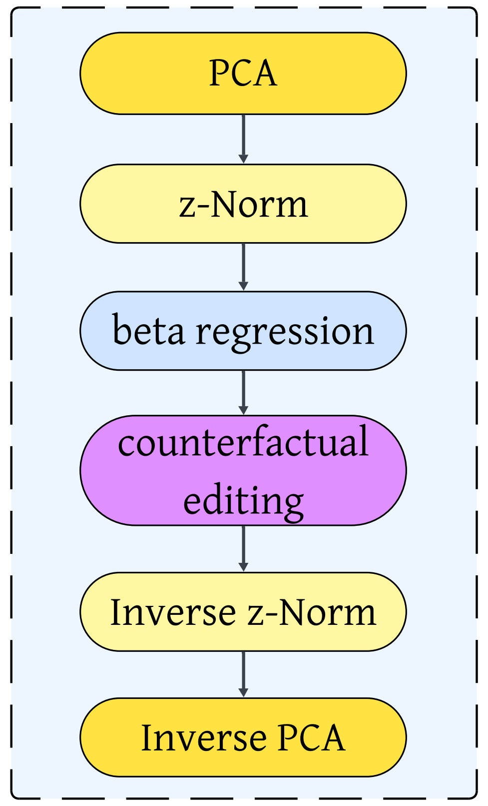


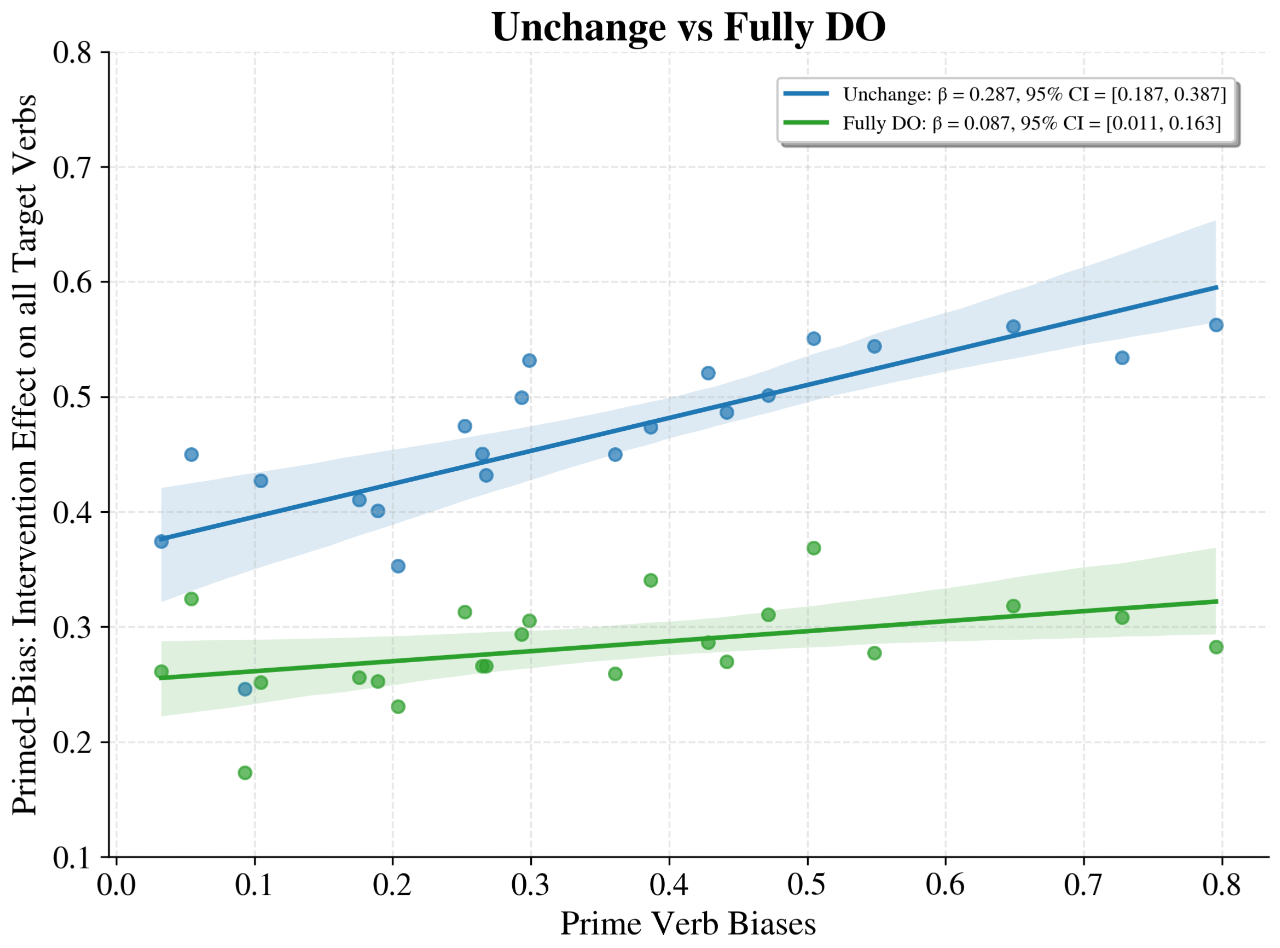
Causal Verification

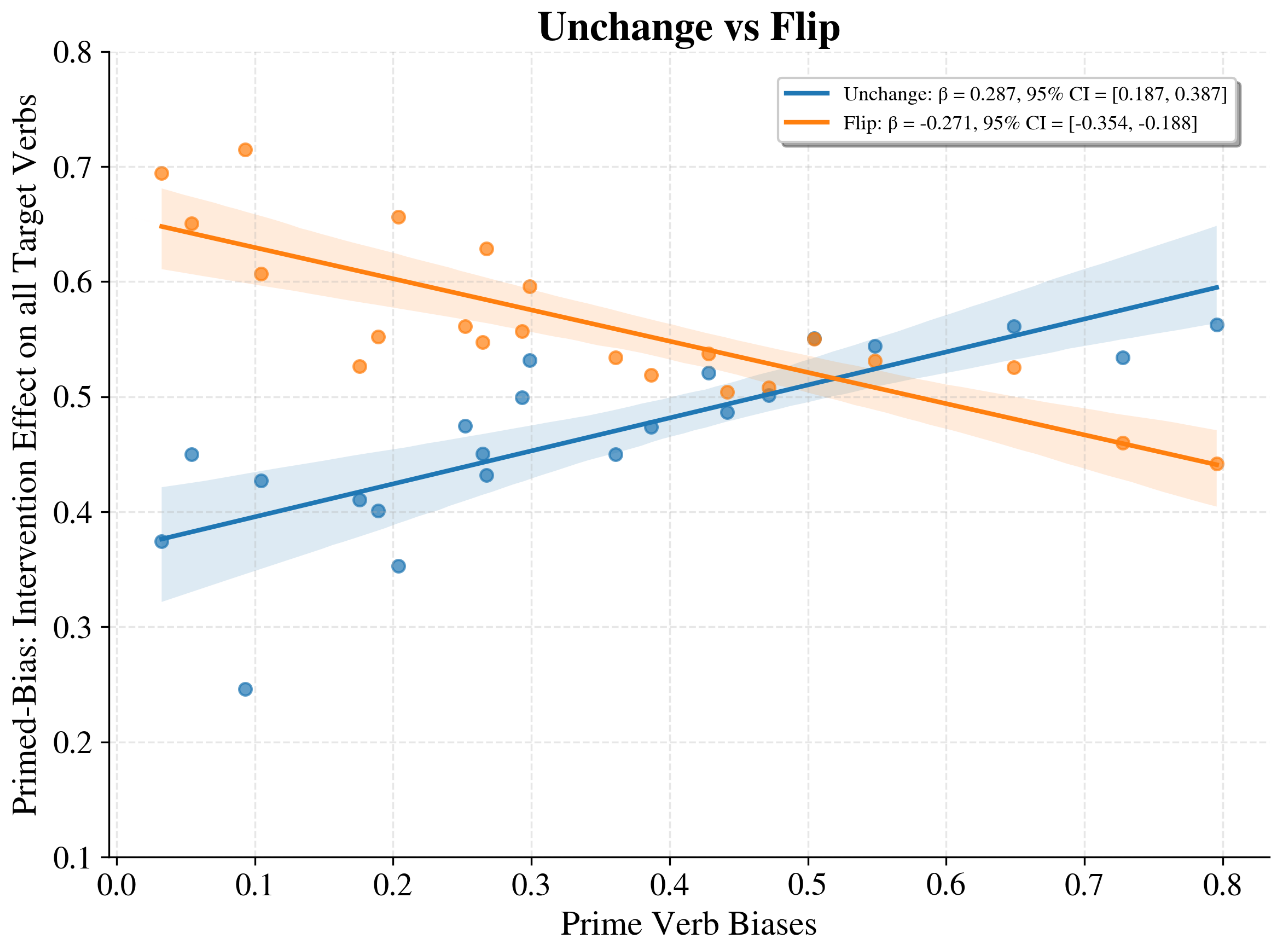




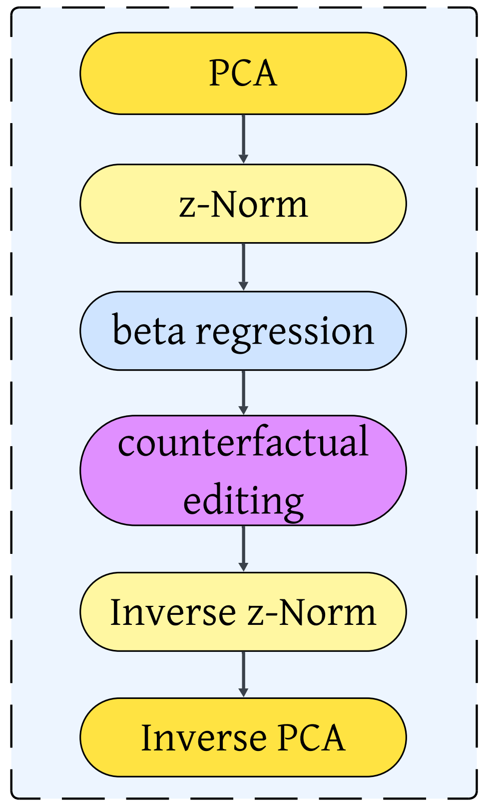
Causal Verification

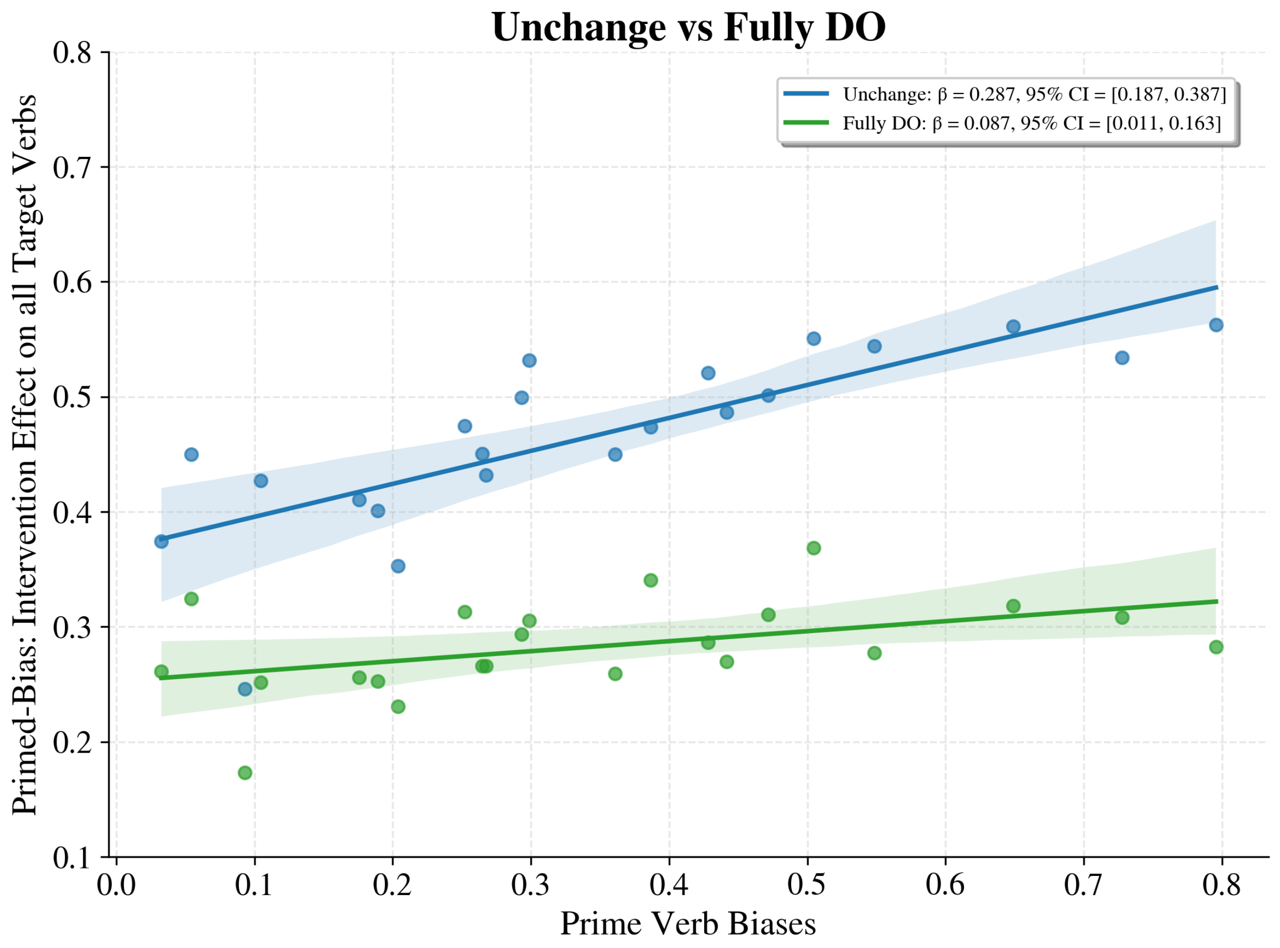


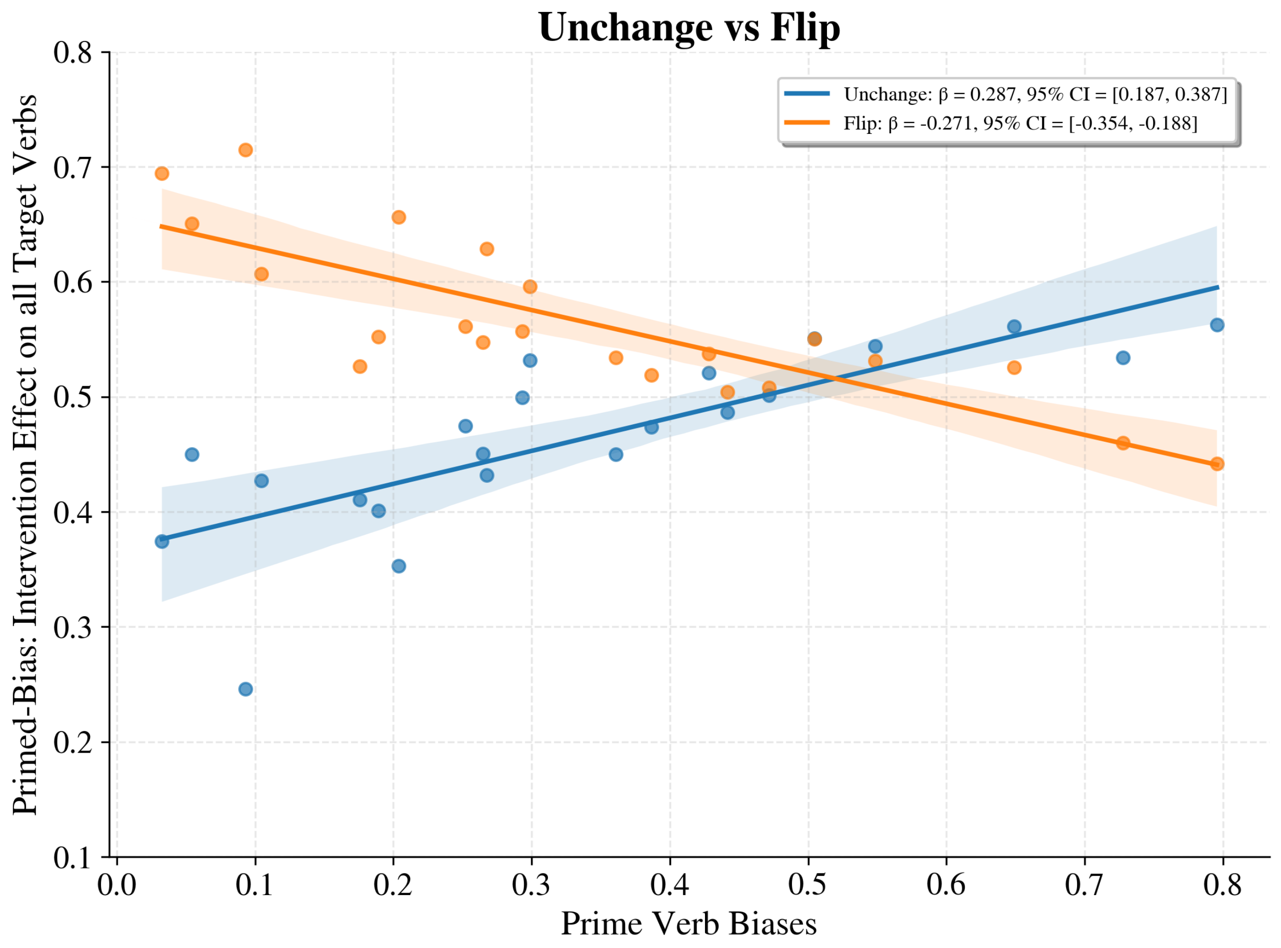


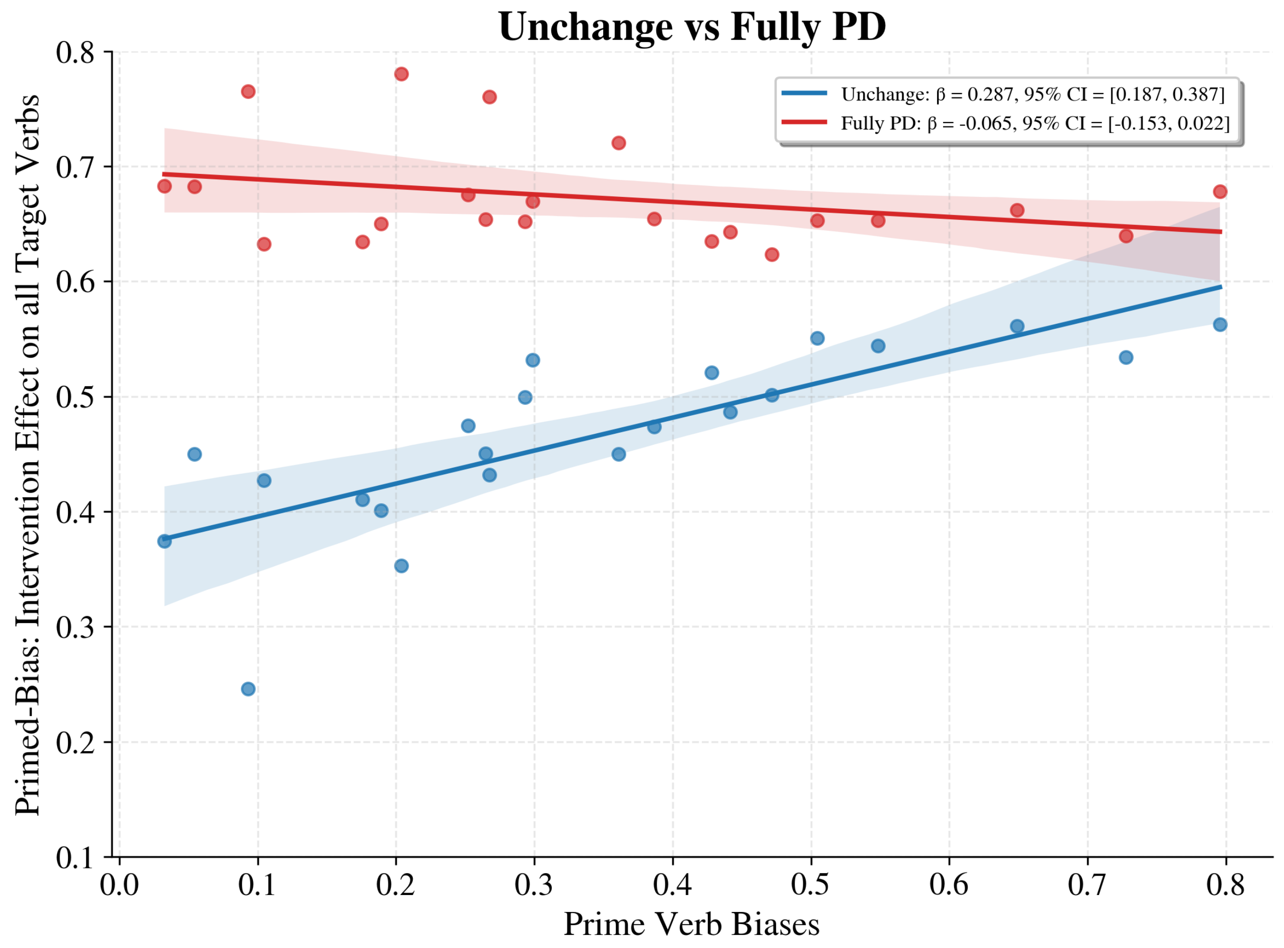


Causal Verification

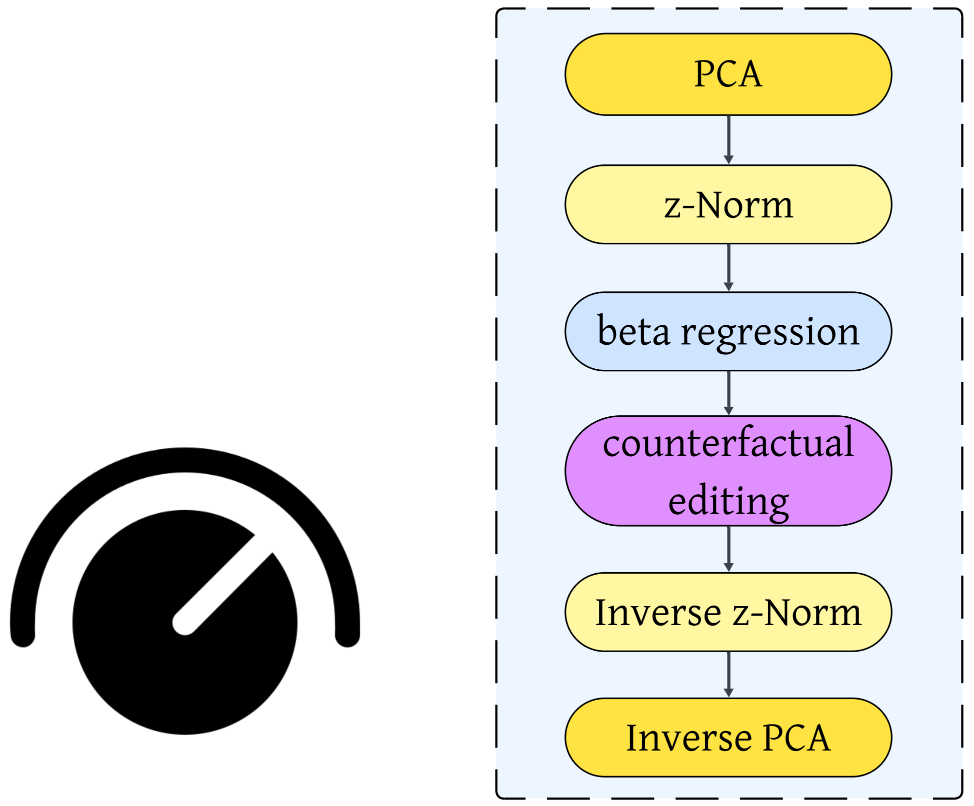


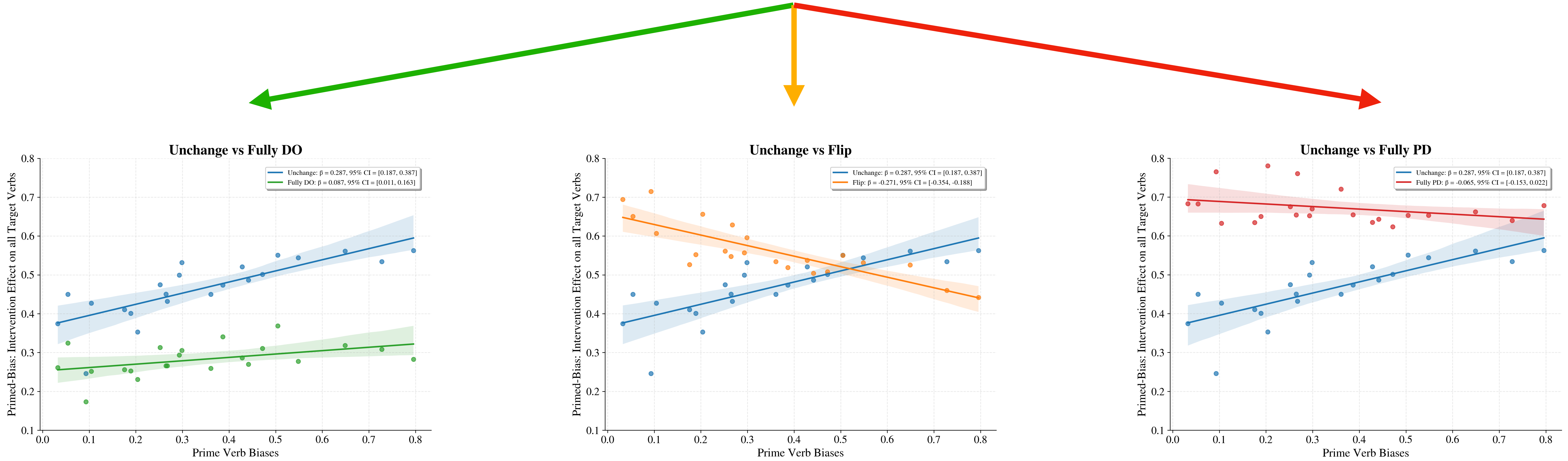






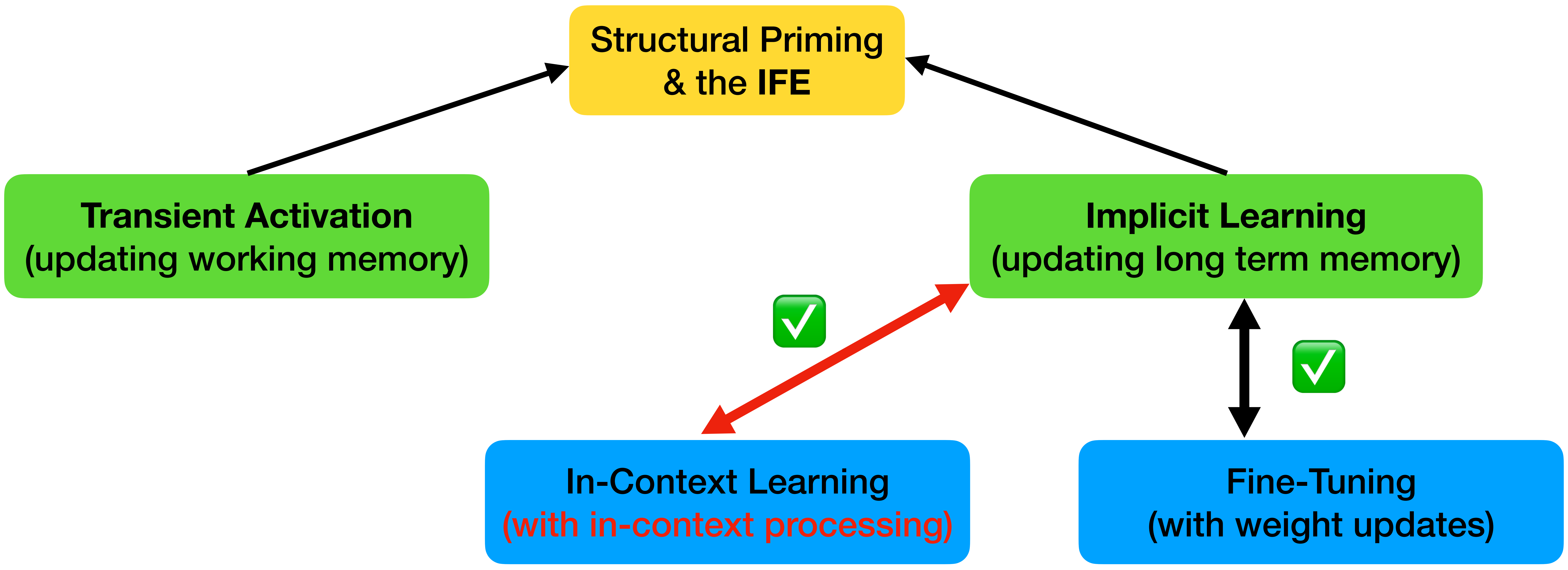
Causal Verification



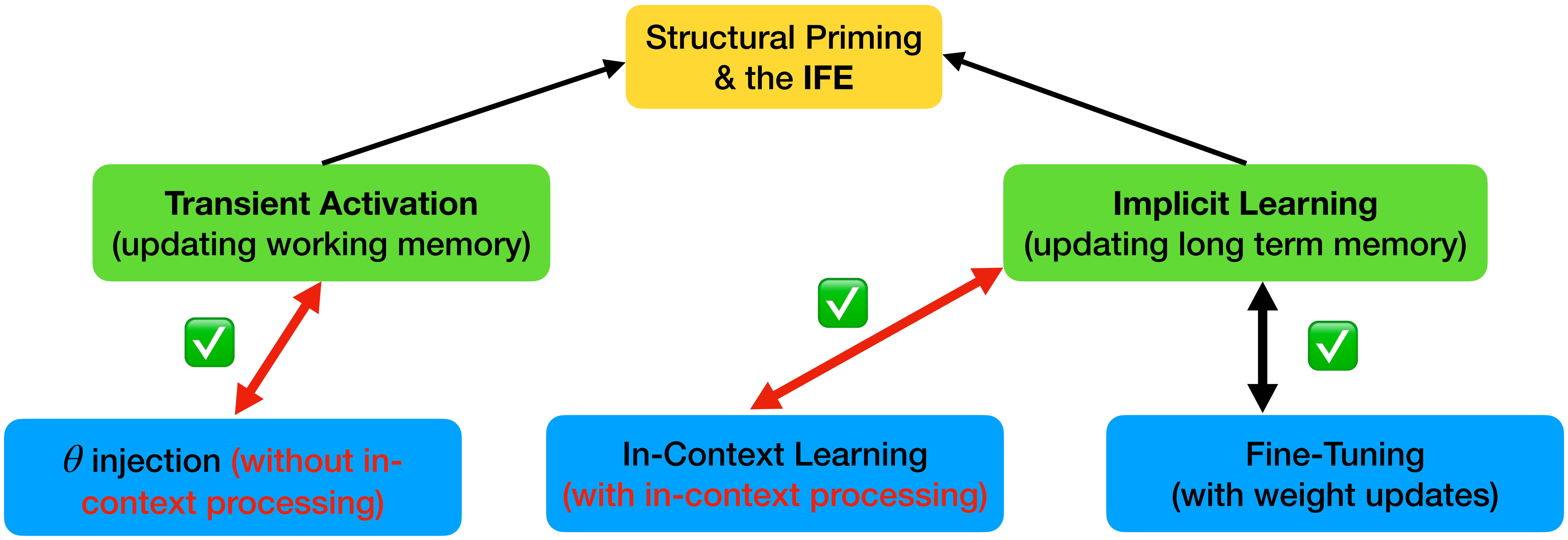


We showed that manipulating the verb biases causes predicted changes in the priming strength;
⇒ The verb bias representation is causally involved in LLM's next-word prediction!

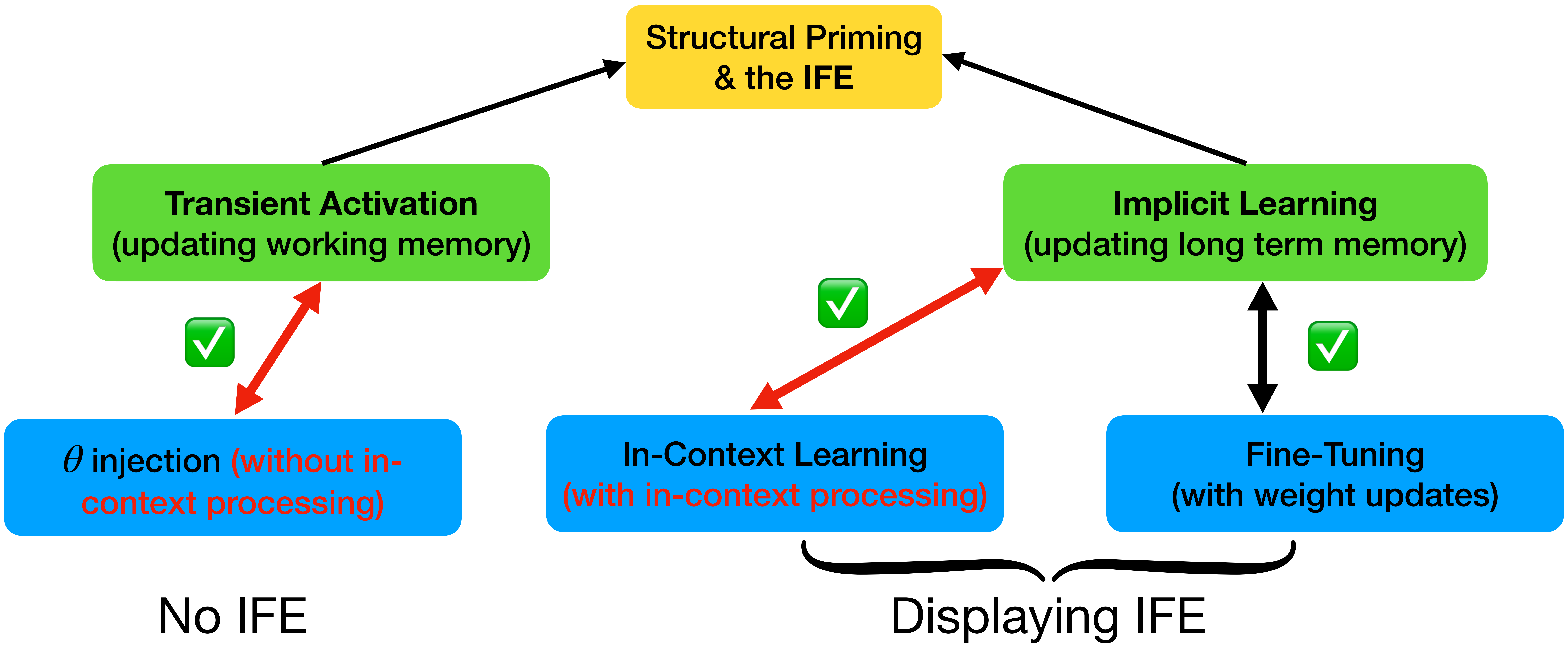
Takeaways from Experiment 2



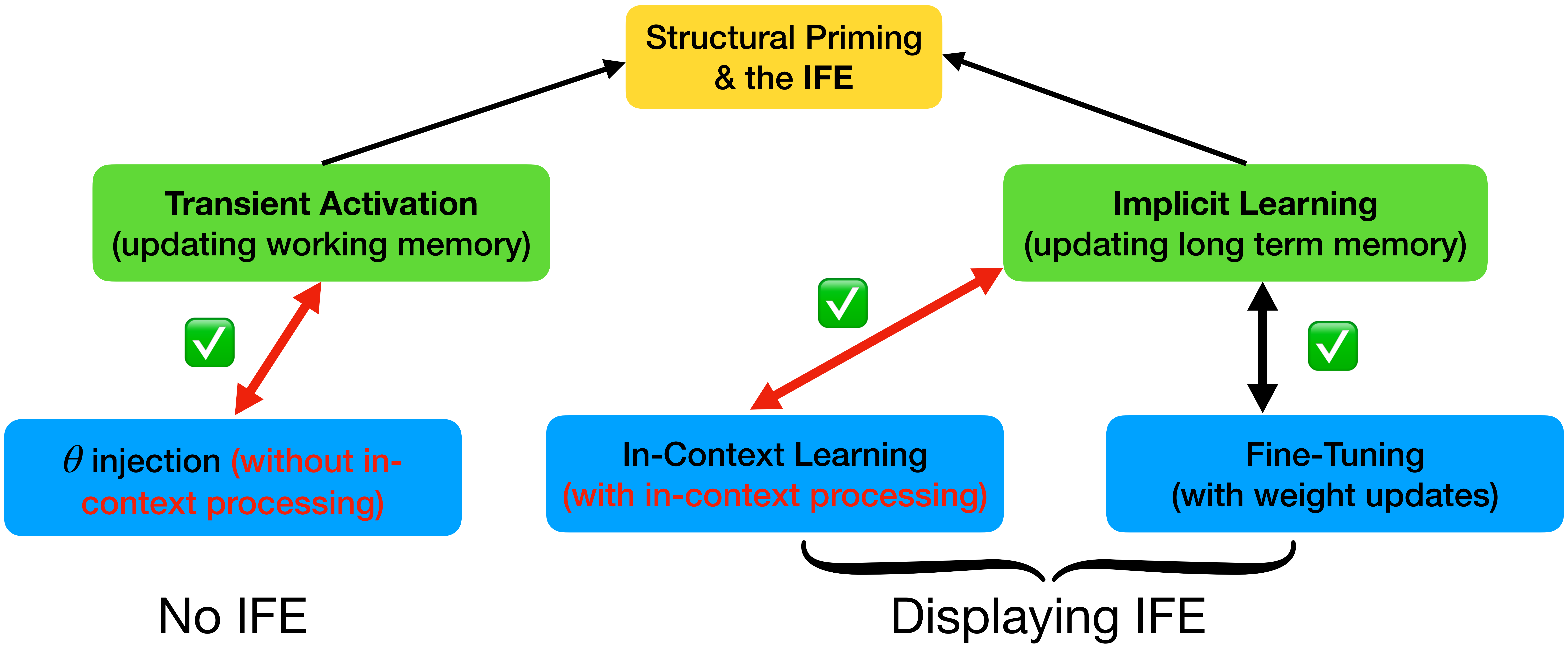
Takeaways from Experiment 2



Takeaways from Experiment 2



Takeaways from Experiment 2



What's crucial for LLMs to adapt to context (in an error-driven way) is to actually have the context window for in-context processing!

Future Steps

Published as a conference paper at COLM 2025

Contextualize-then-Aggregate: Circuits for In-Context Learning in Gemma-2 2B

Aleksandra Bakalova, Yana Veitsman, Xinting Huang & Michael Hahn
Saarland Informatics Campus, Saarland University
{abakalov, yanav, xhuang, mhahn}@lst.uni-saarland.de

Future Steps

Published as a conference paper at COLM 2025

Contextualize-then-Aggregate: Circuits for In-Context Learning in Gemma-2 2B

Aleksandra Bakalova, Yana Veitsman, Xinting Huang & Michael Hahn
Saarland Informatics Campus, Saarland University
{abakalov, yanav, xhuang, mhahn}@lst.uni-saarland.de

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!

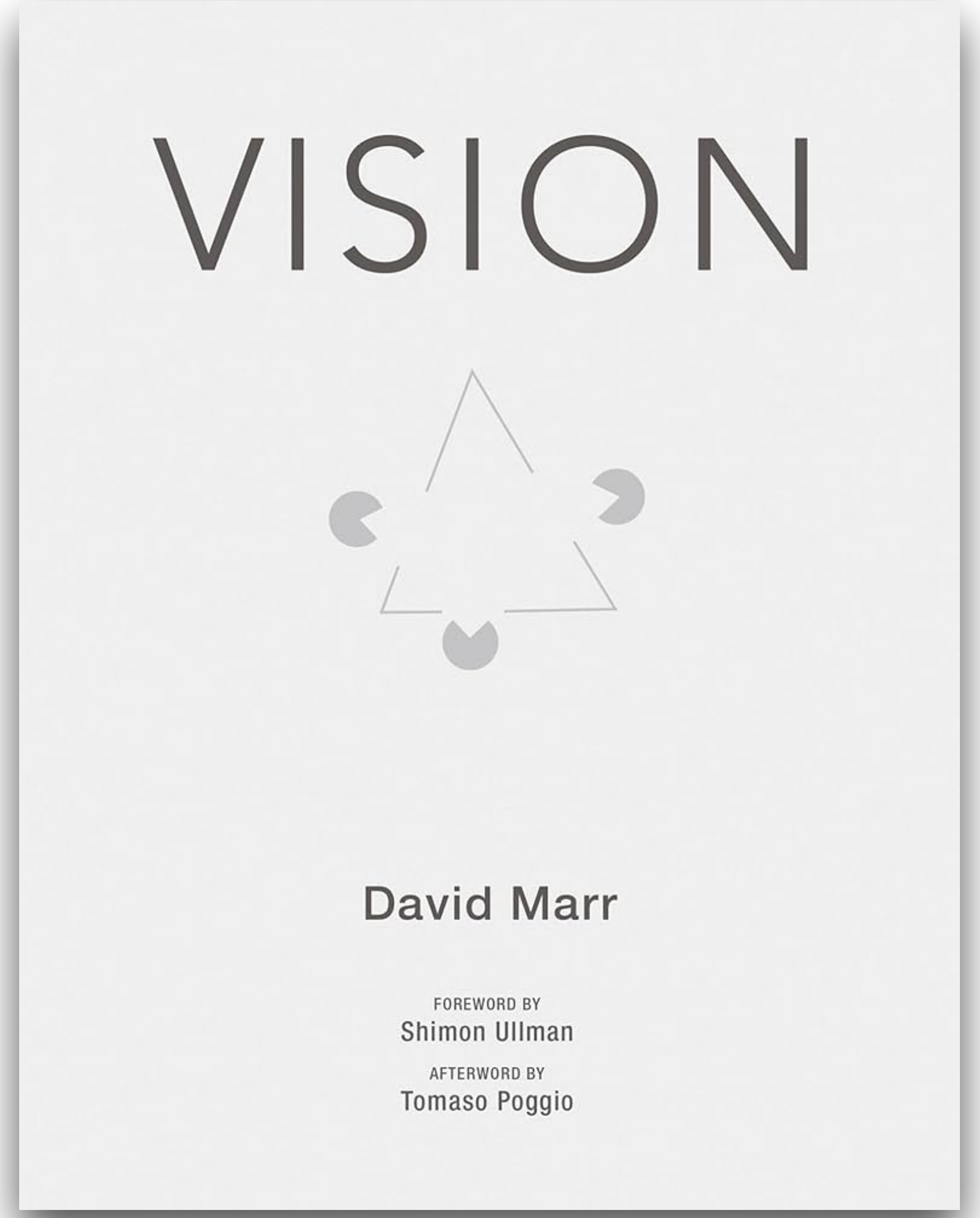
Q&A Discussion

Appendices

Revisiting Marr’s Levels of Analysis

Multiple facets for studying information processing systems

[Marr, 1982]

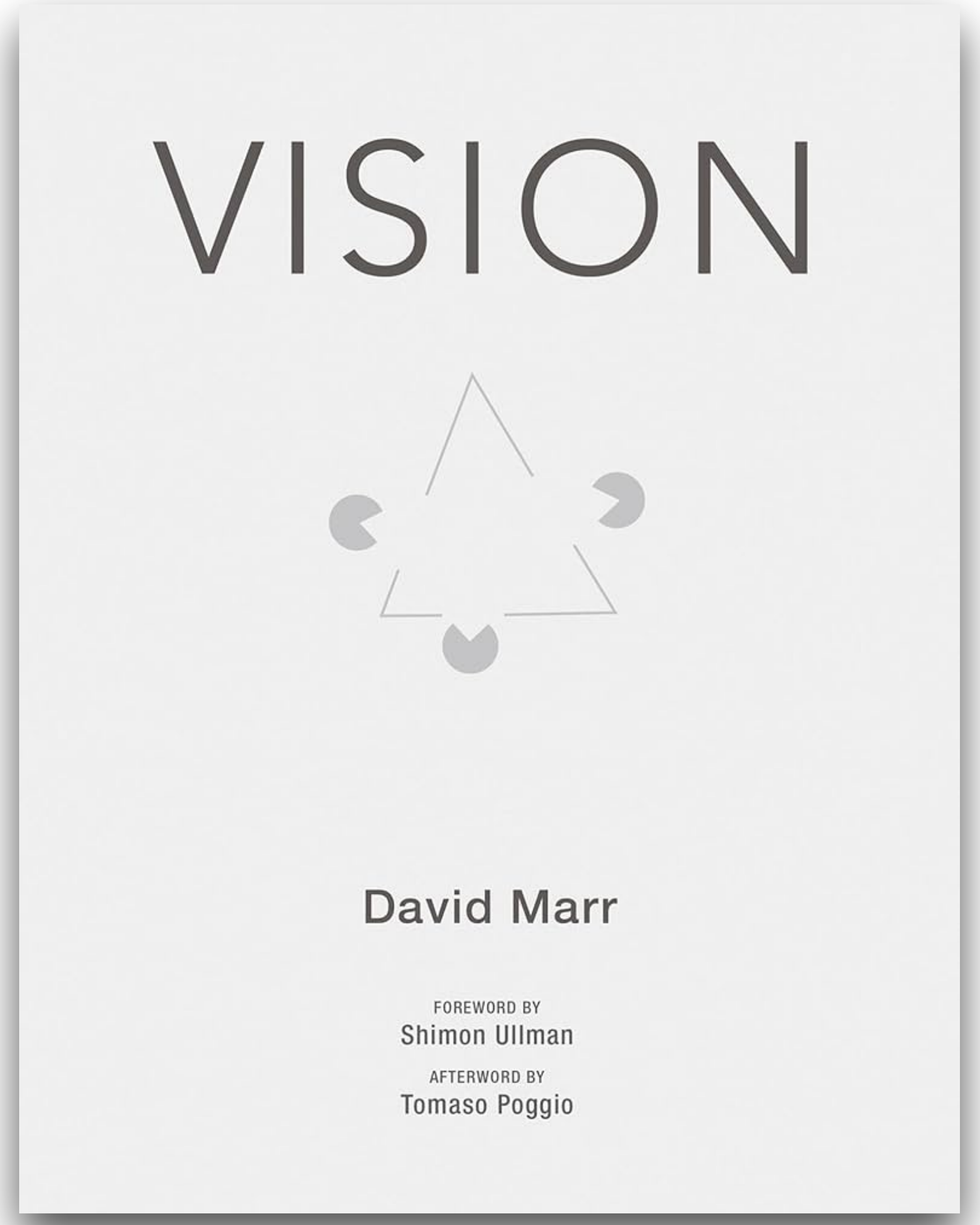


Revisiting Marr's Levels of Analysis

Multiple facets for studying information processing systems

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

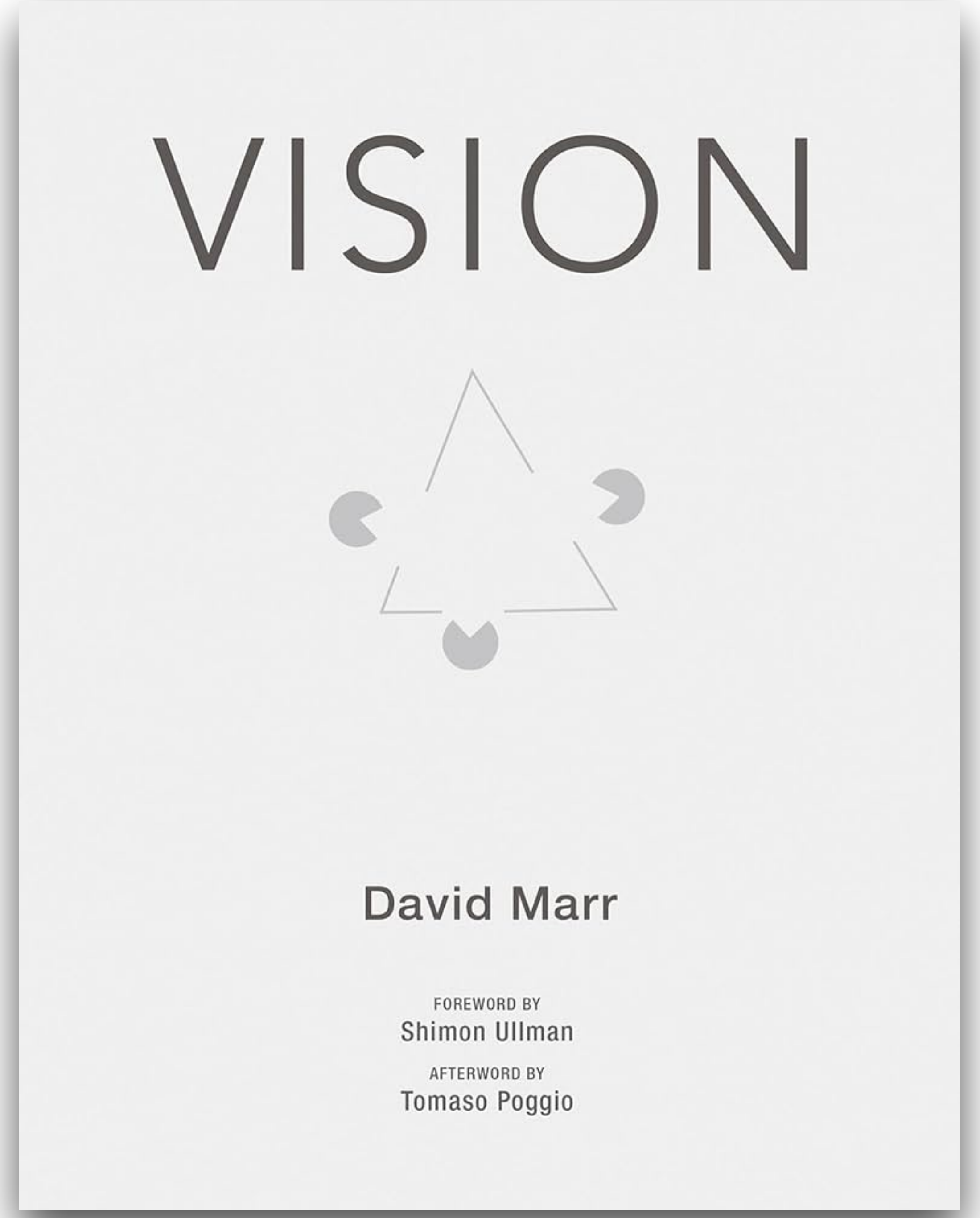
[Marr, 1982]



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*



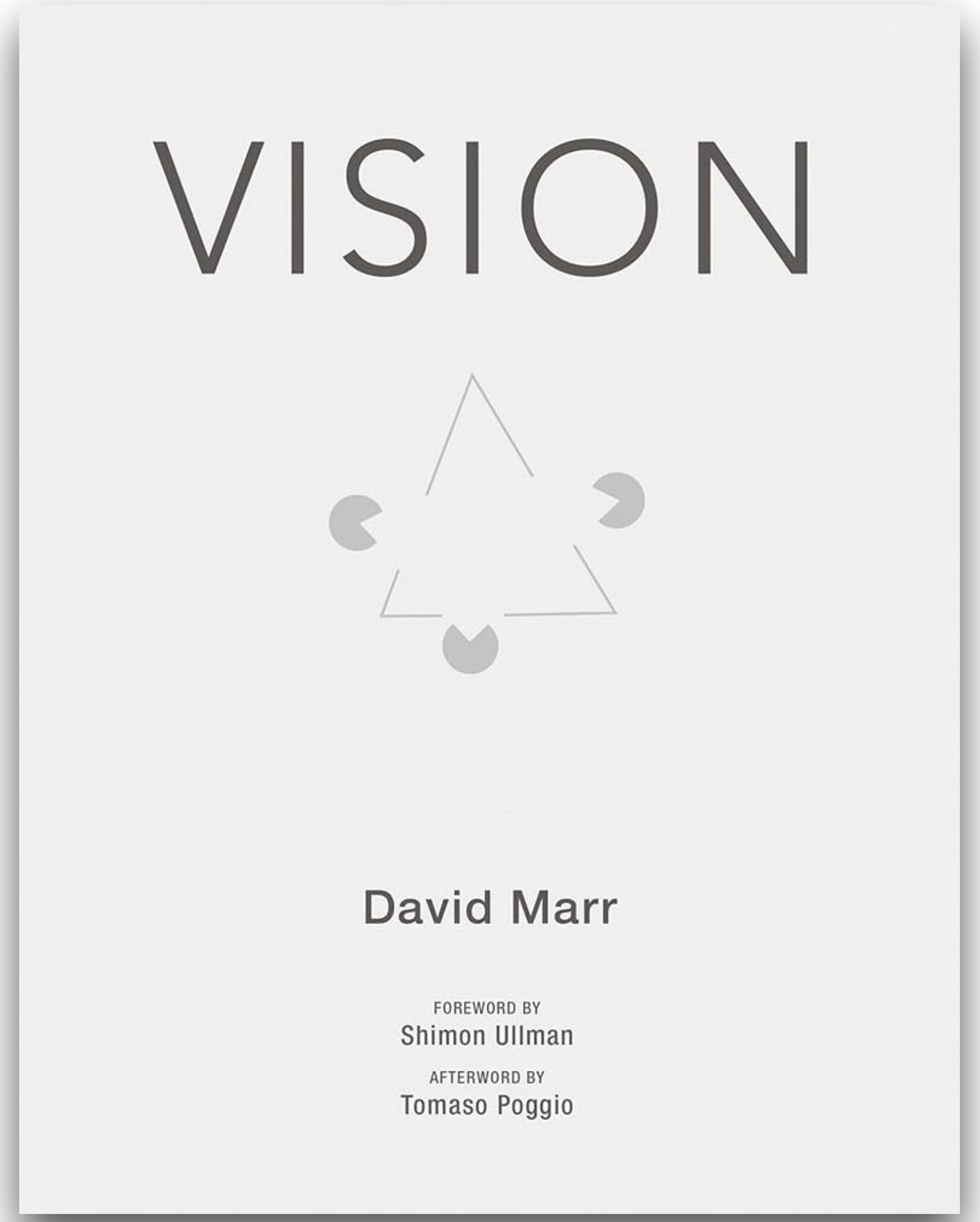
[Marr, 1982]

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*

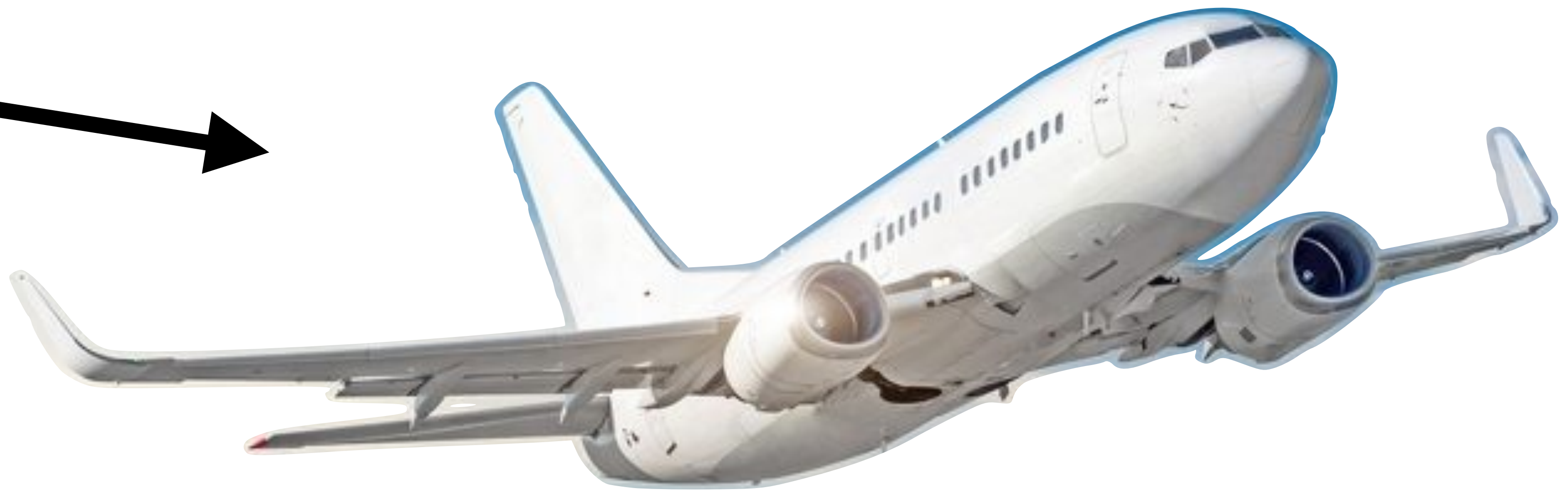
[Marr, 1982]

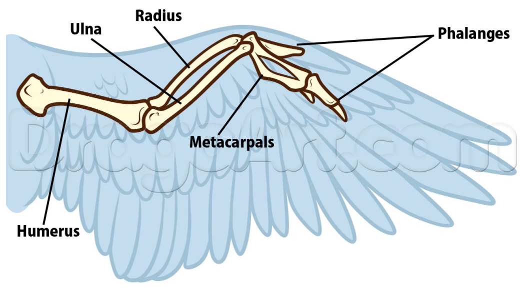


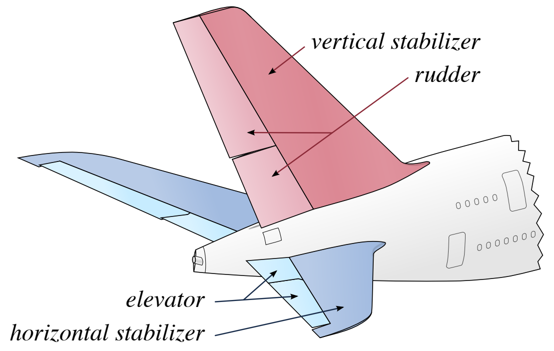
Revisiting Birds vs. Airplanes

Does understanding airplanes inform us how birds fly?

Be able to fly



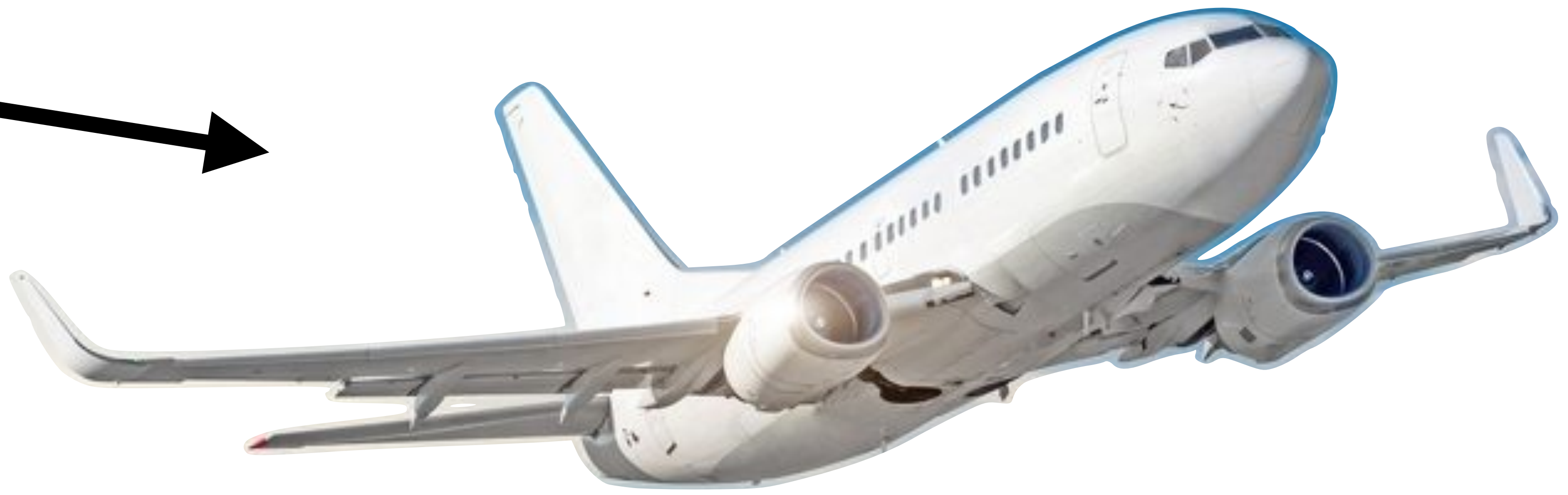




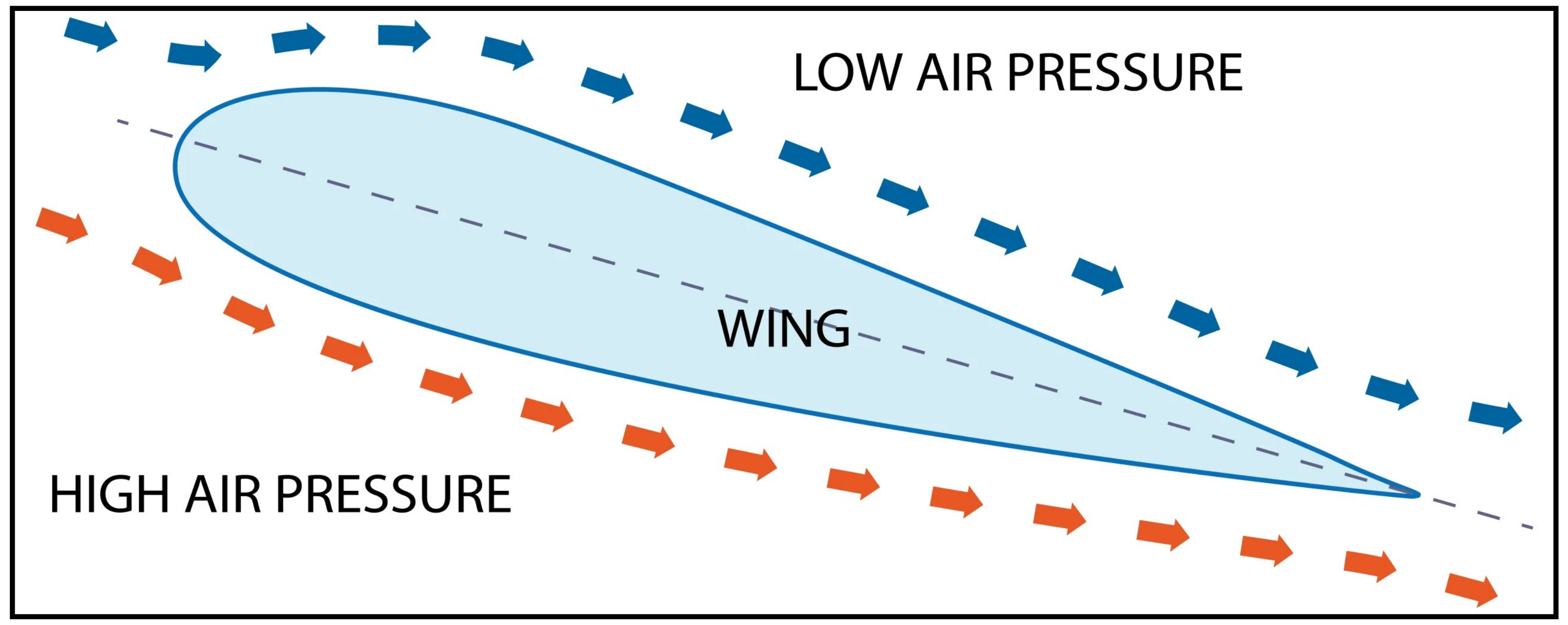
Revisiting Birds vs. Airplanes

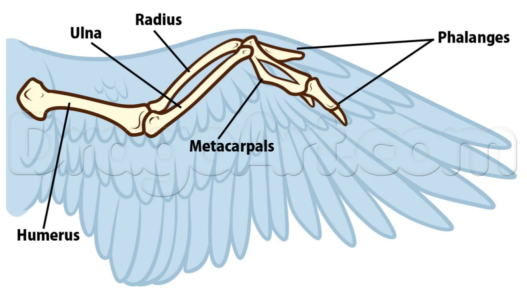
Does understanding airplanes inform us how birds fly?

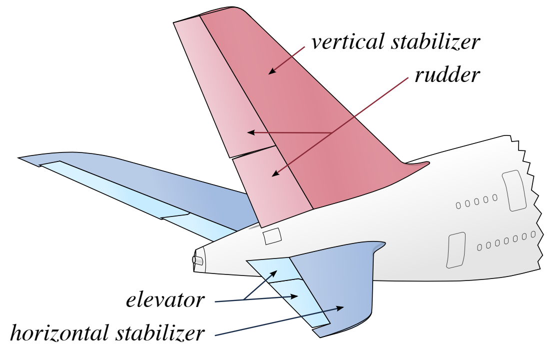
Be able to fly



Following aerodynamics



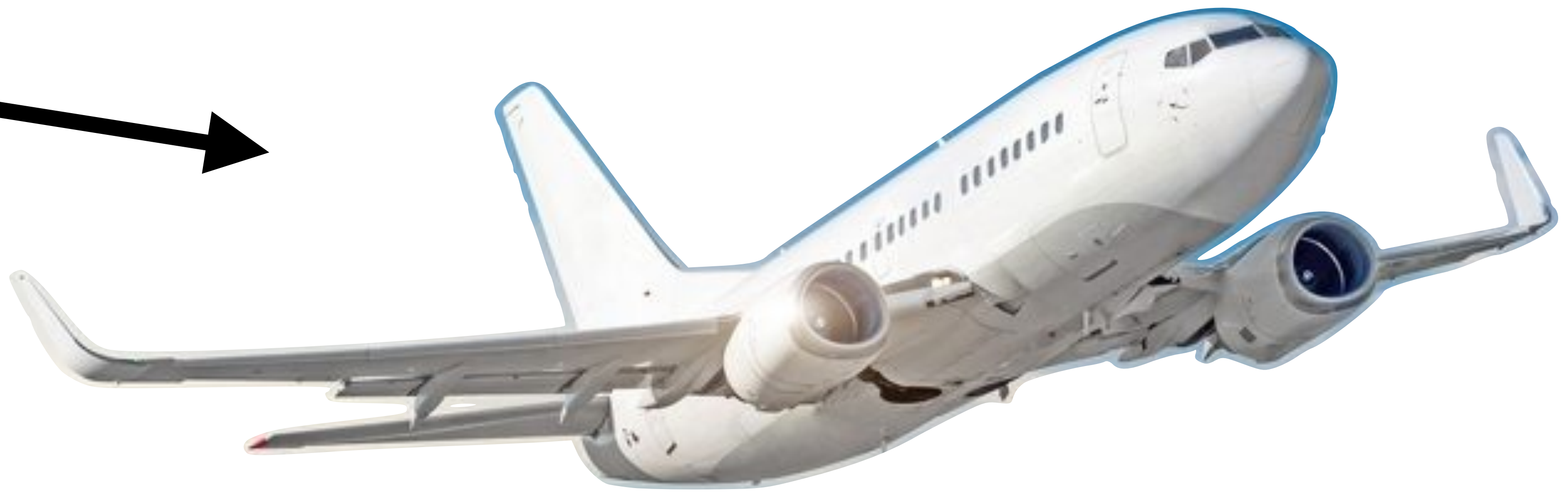




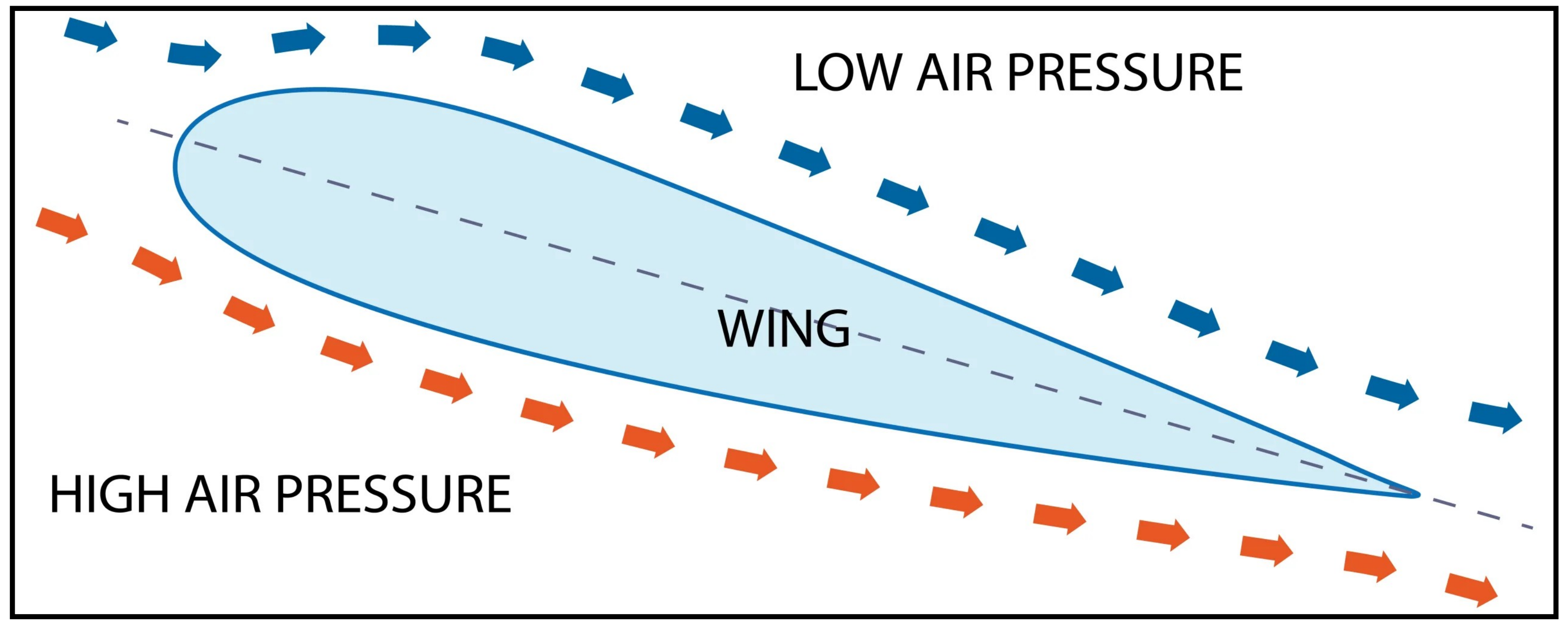
Revisiting Birds vs. Airplanes

Does understanding airplanes inform us how birds fly?

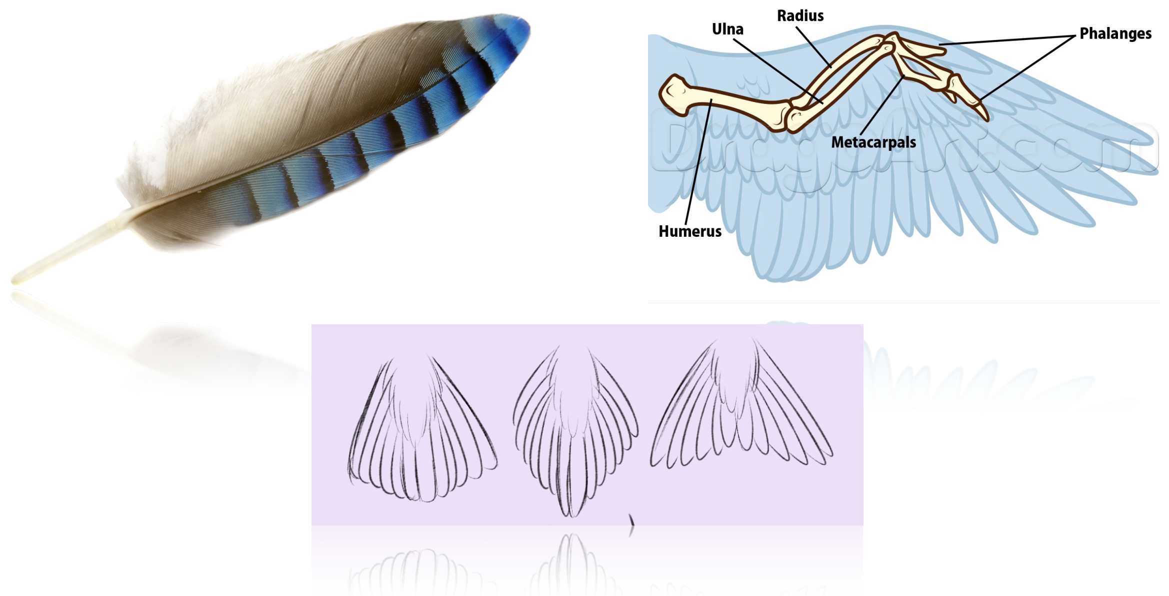
Be able to fly

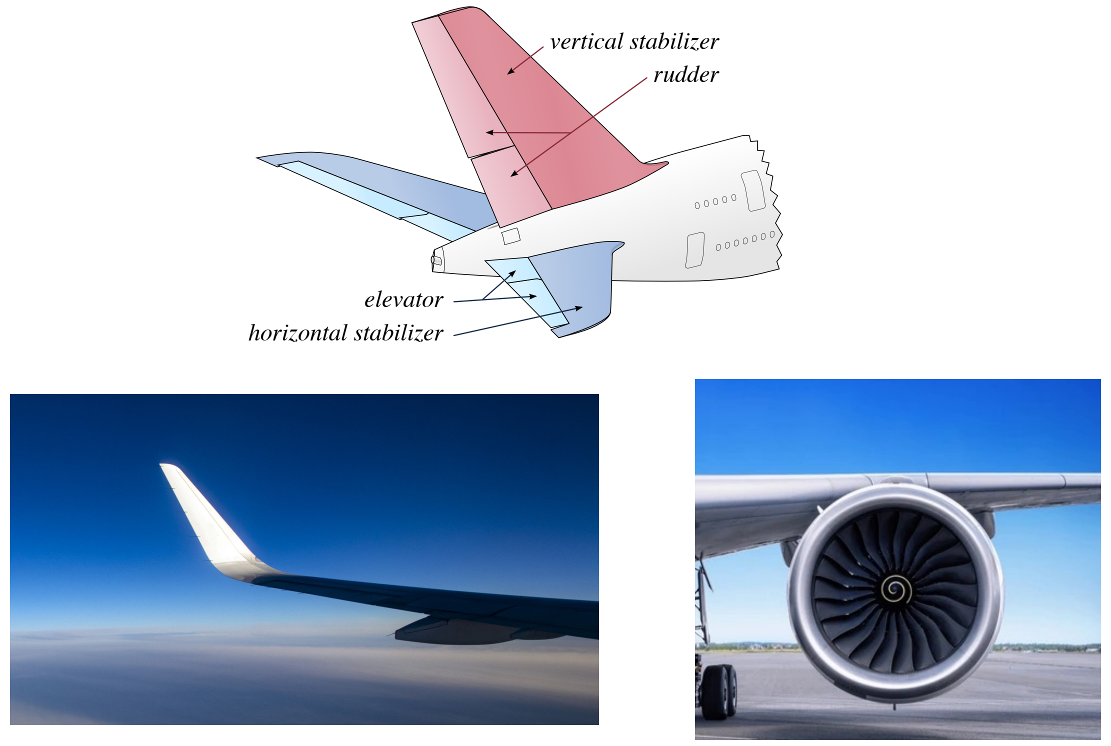


Following aerodynamics



A general "algorithm" of flying spelled out in aerodynamic terms

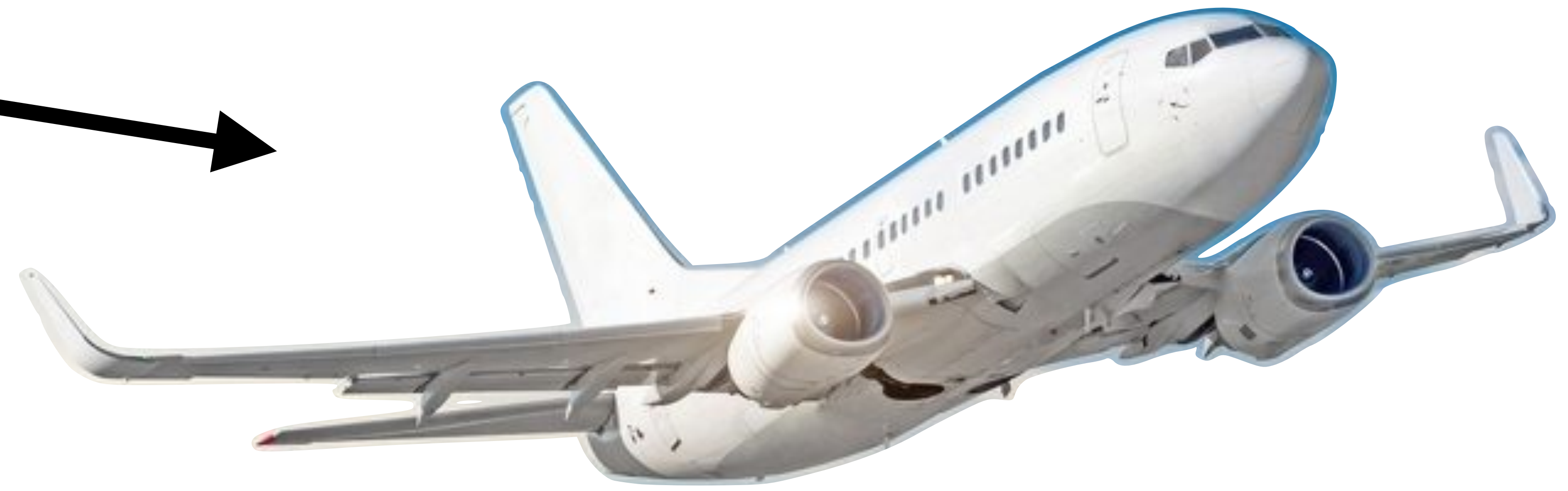




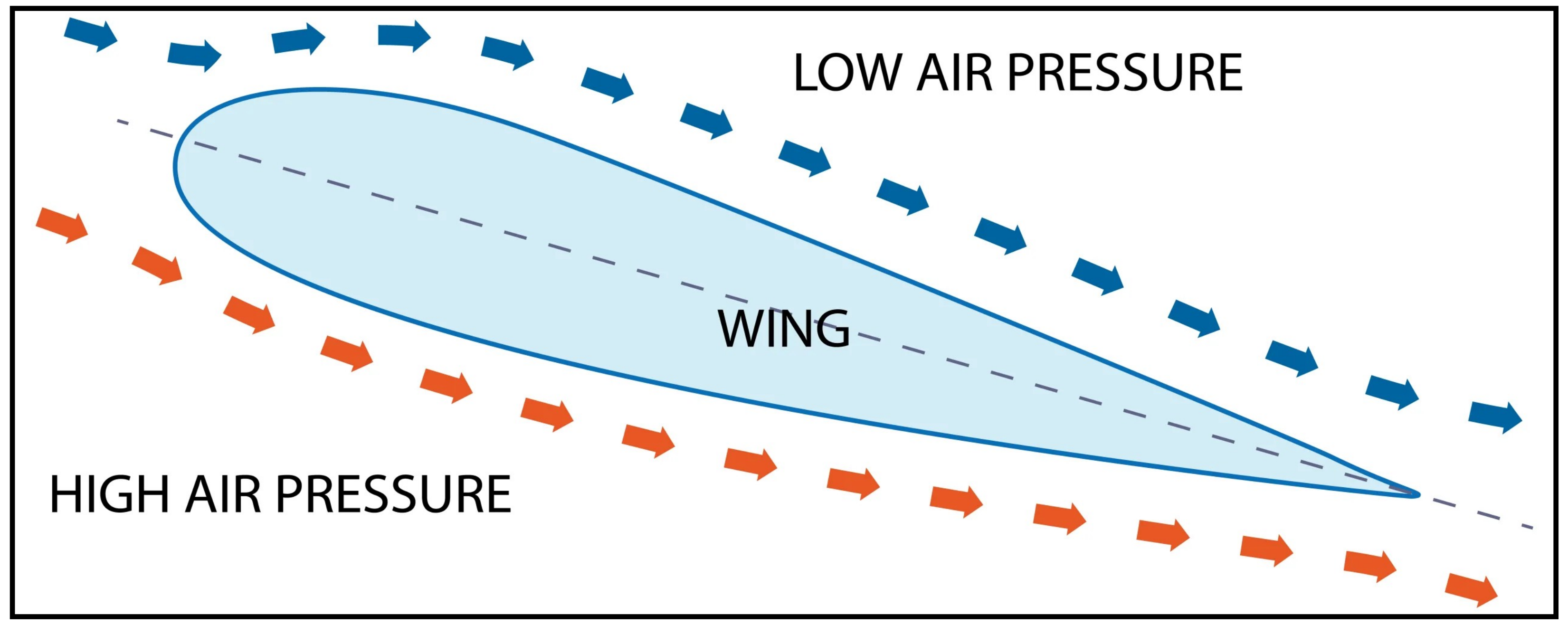
Revisiting Birds vs. Airplanes

Does understanding airplanes inform us how birds fly?

Be able to fly

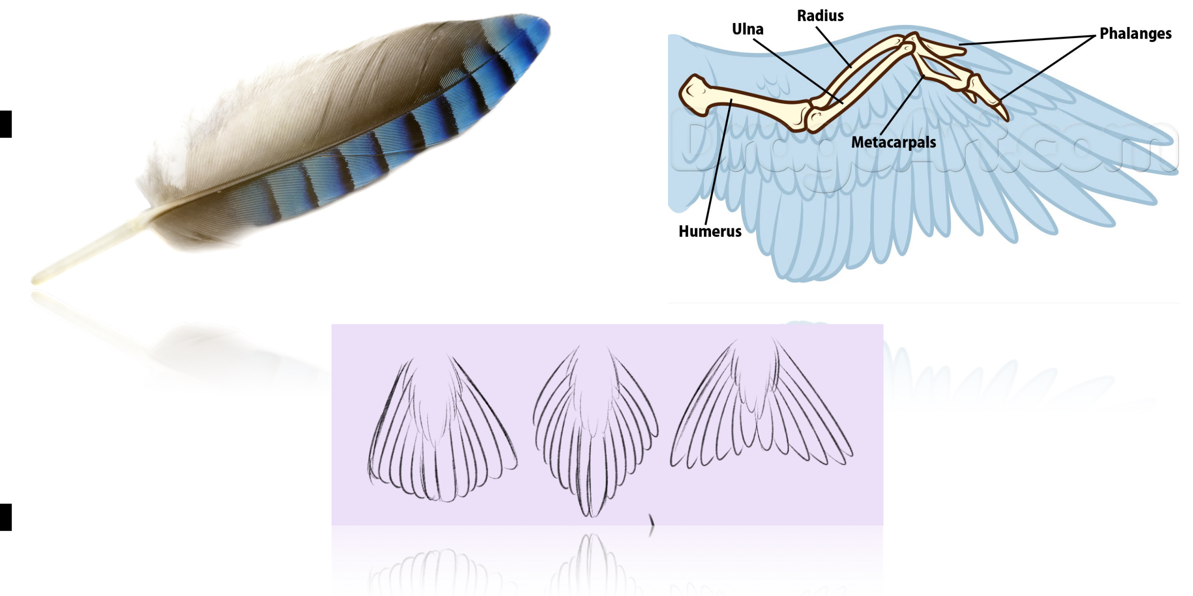


Following aerodynamics

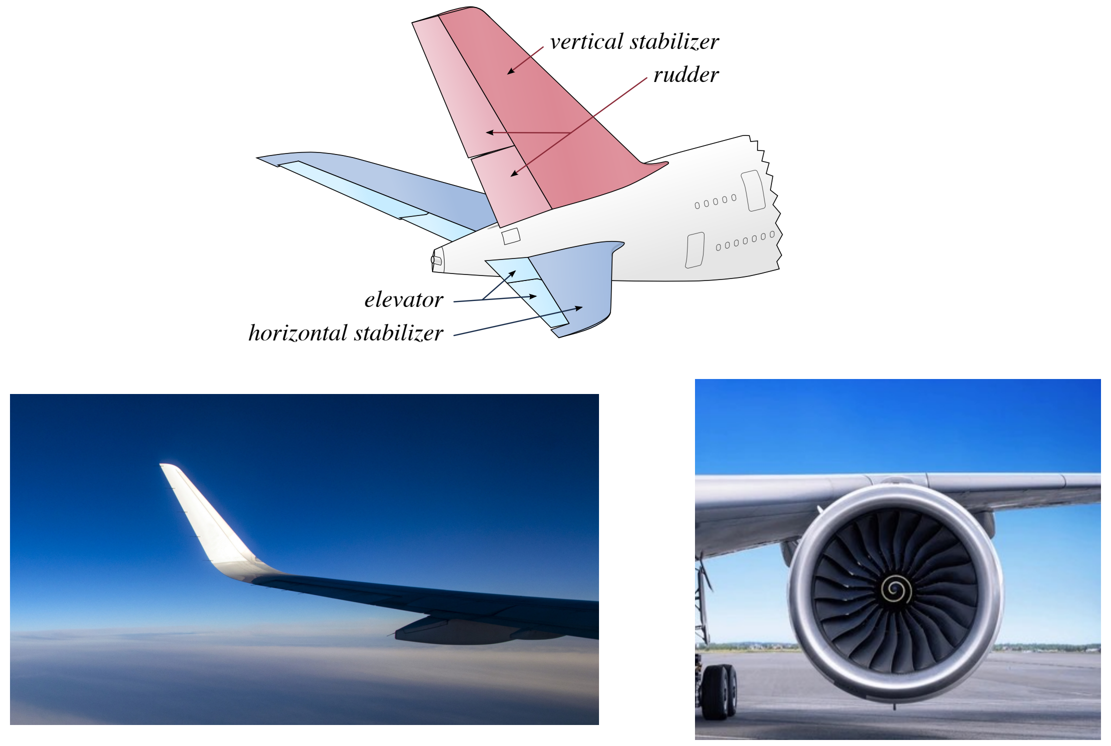


Functionally approximating

A general "algorithm" of flying spelled out in aerodynamic terms

{ 

}

{ 

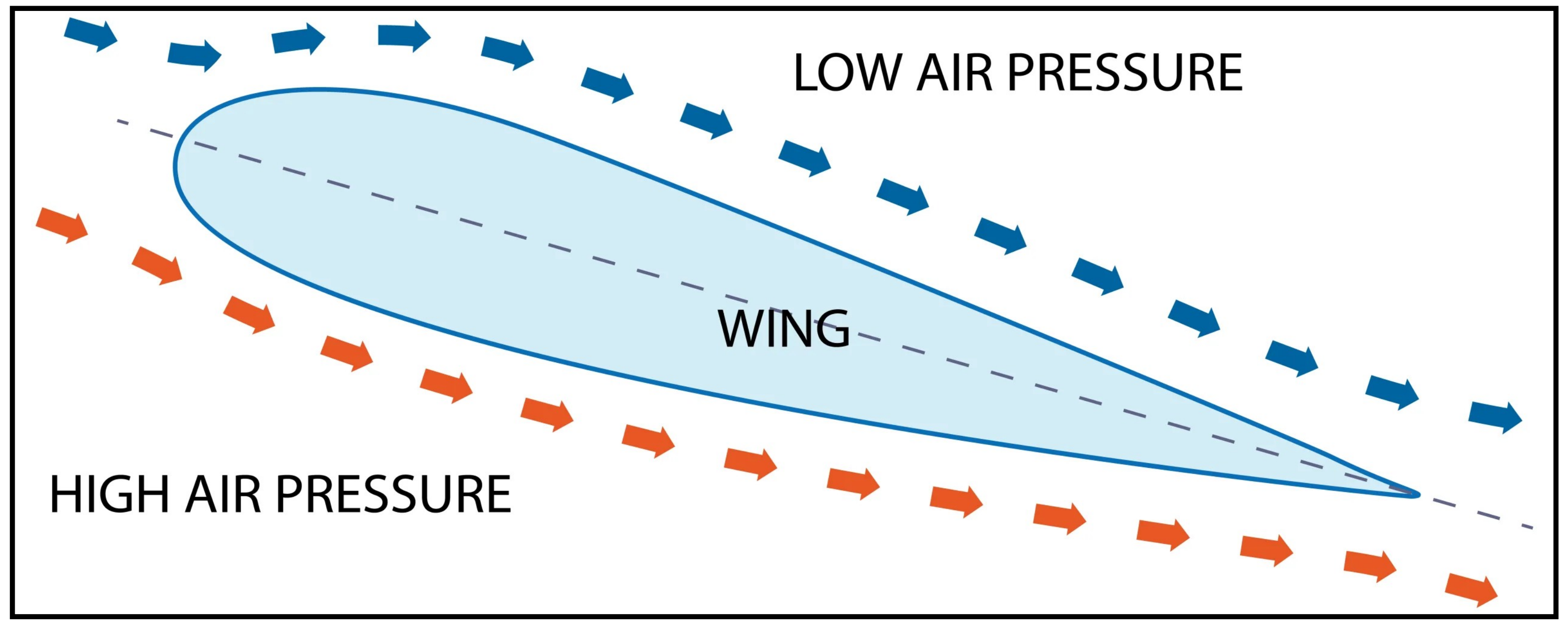
}

Revisiting Birds vs. Airplanes

Does understanding airplanes inform us how birds fly?

Be able to fly

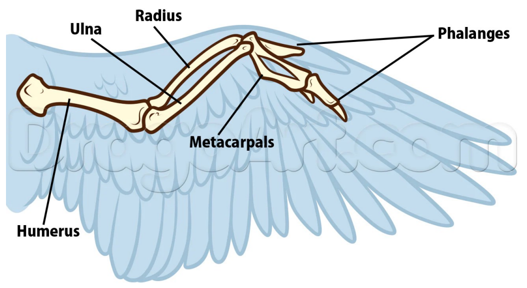
Following aerodynamics



A general "algorithm" of flying spelled out in aerodynamic terms

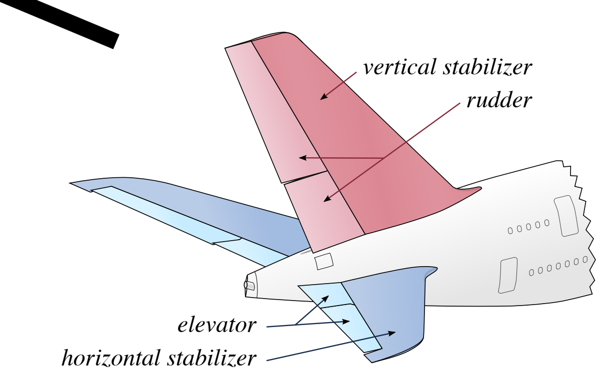
Functionally approximating

{




}

Functionally approximating

{




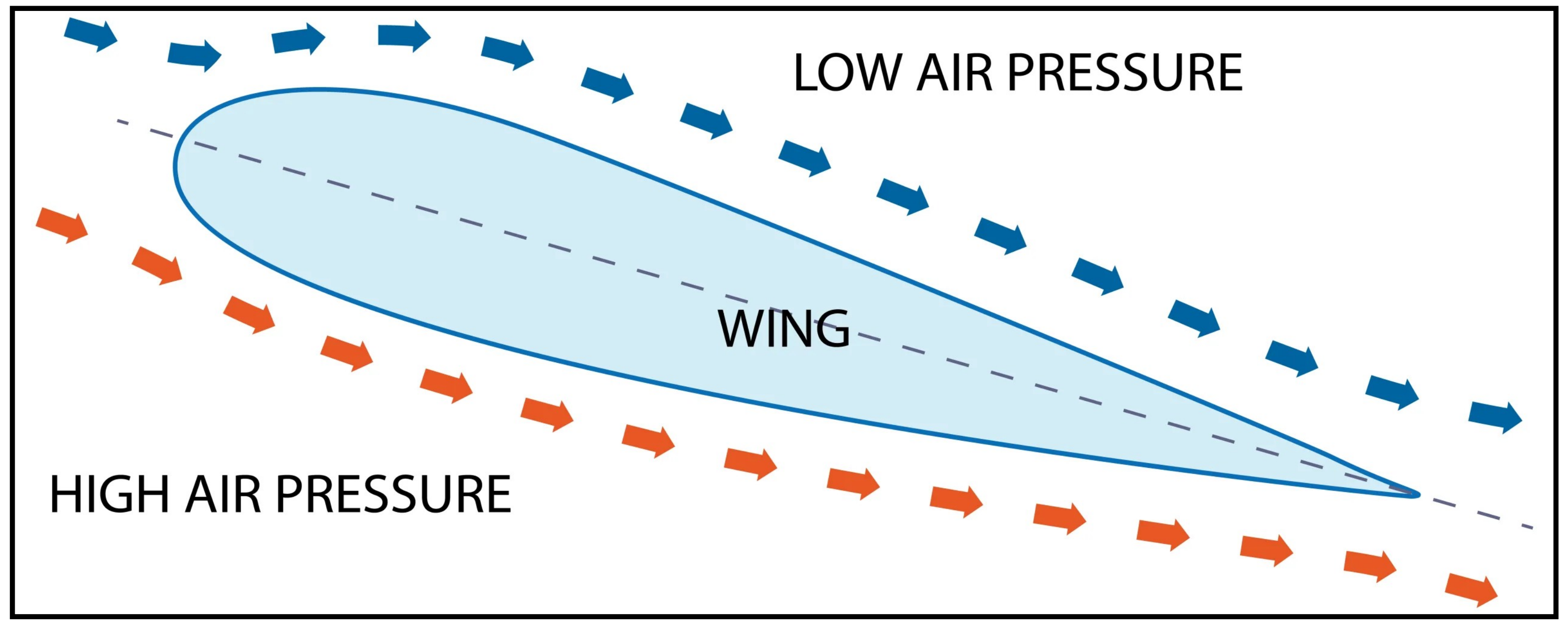
}

Revisiting Birds vs. Airplanes

Does understanding airplanes inform us how birds fly?

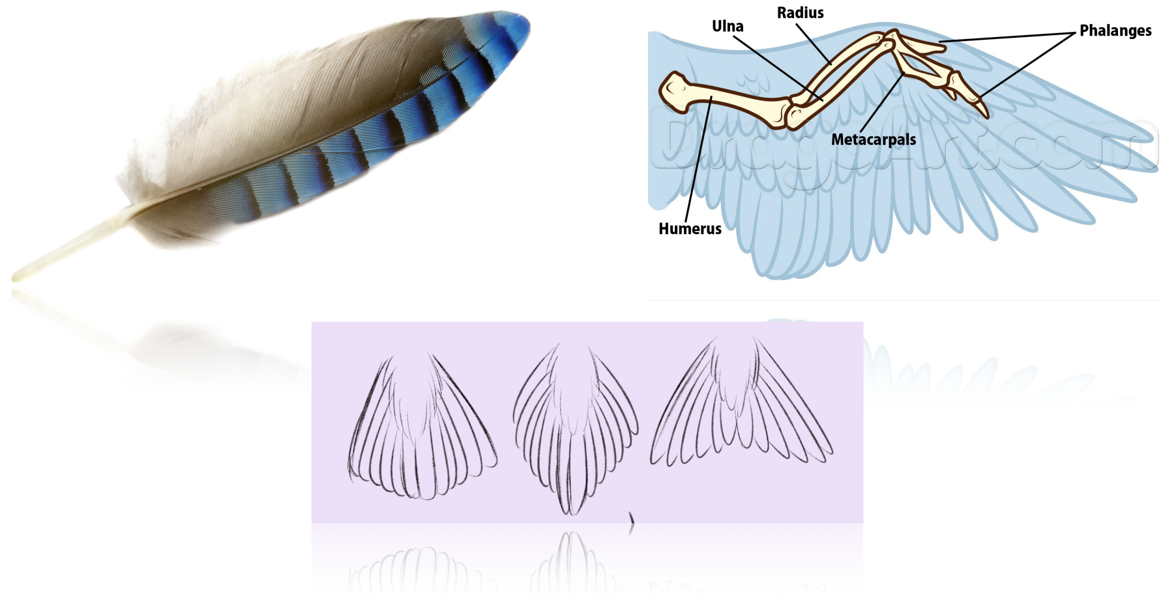
Be able to fly

Following aerodynamics

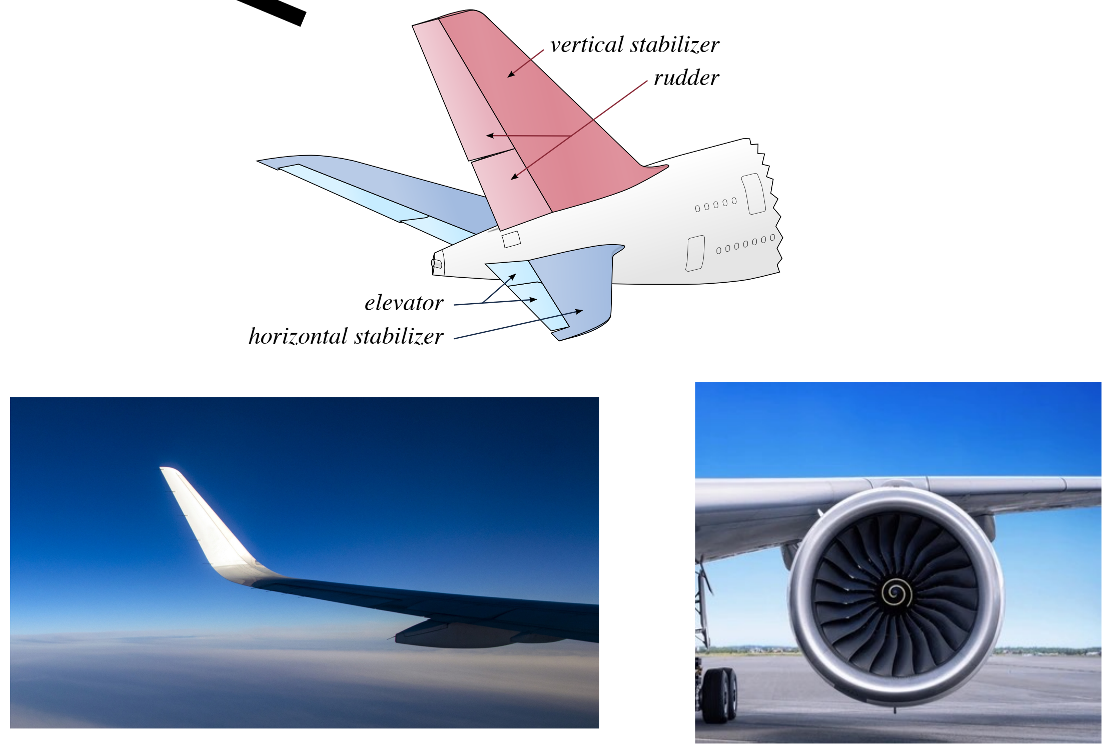


A general "algorithm" of flying spelled out in aerodynamic terms

Functionally approximating



Functionally approximating



Current Proposal: The Full Picture

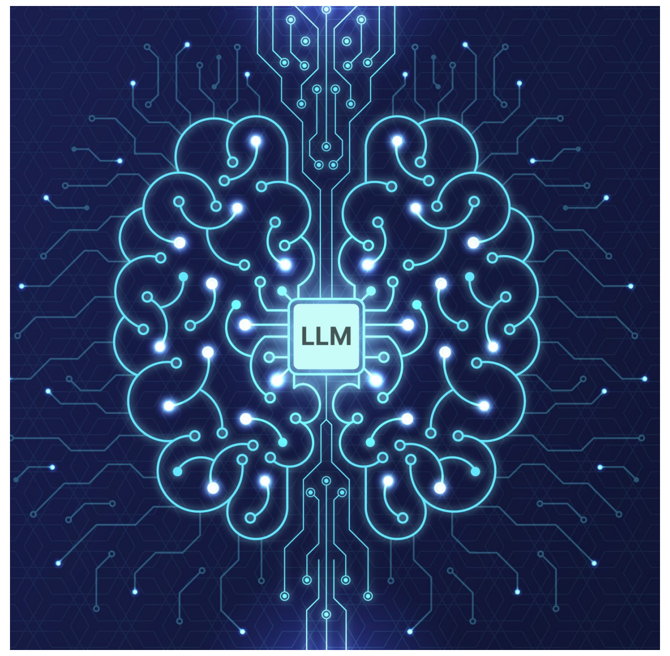
Current Proposal: The Full Picture

Computational Level

Current Proposal: The Full Picture

Computational Level

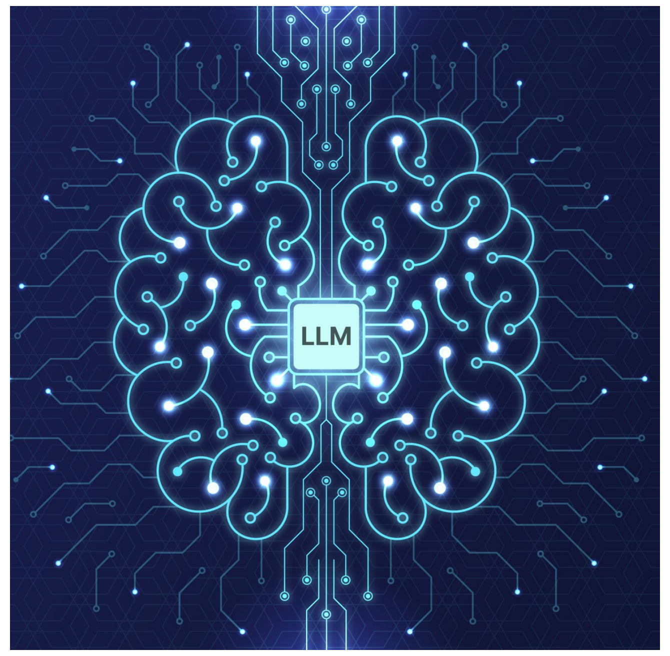
Language **in the broader design of intelligent systems**



Current Proposal: The Full Picture

Language in the broader design of intelligent systems

Computational Level

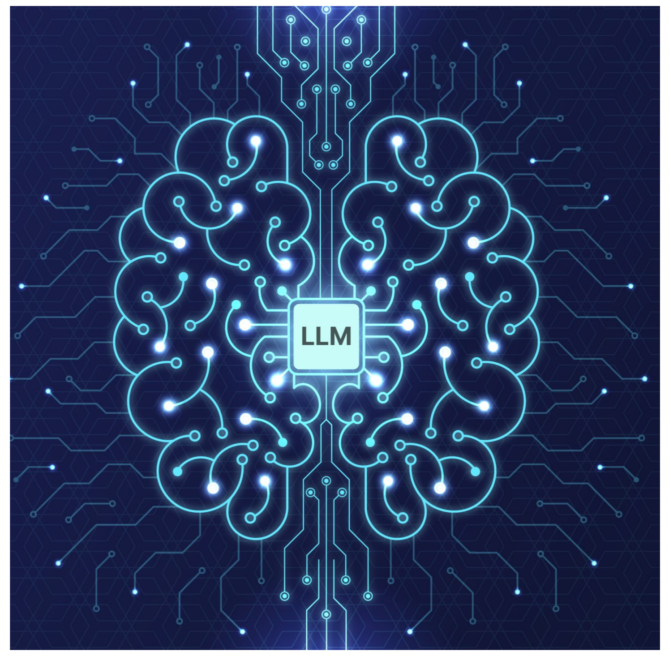


Algorithmic Level ONE

Current Proposal: The Full Picture

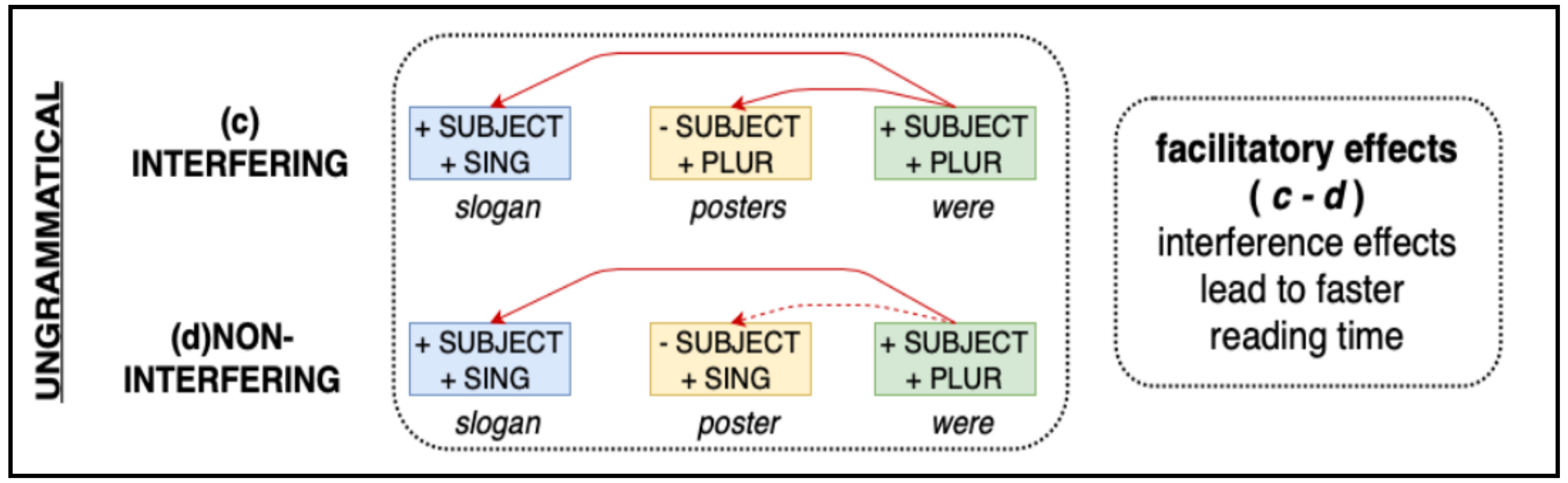
Computational Level

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



Psycholinguistic theories (?)

Algorithmic Level ONE

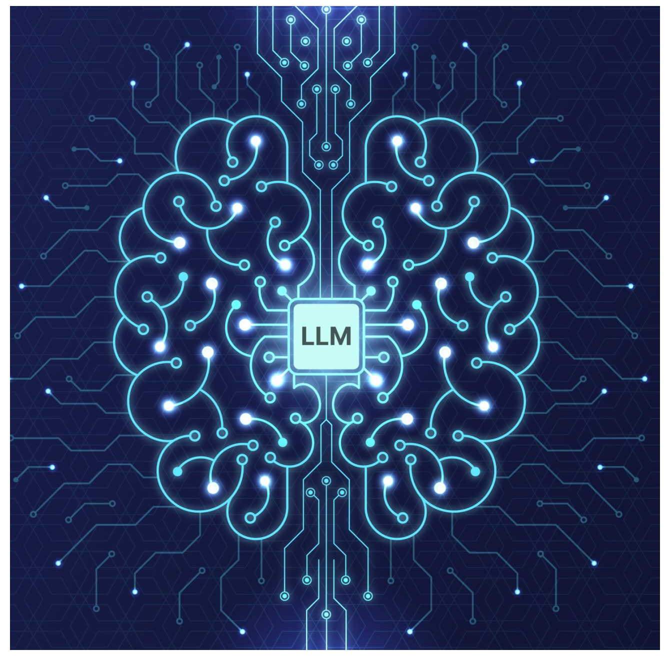


An "algorithm" of sentence processing

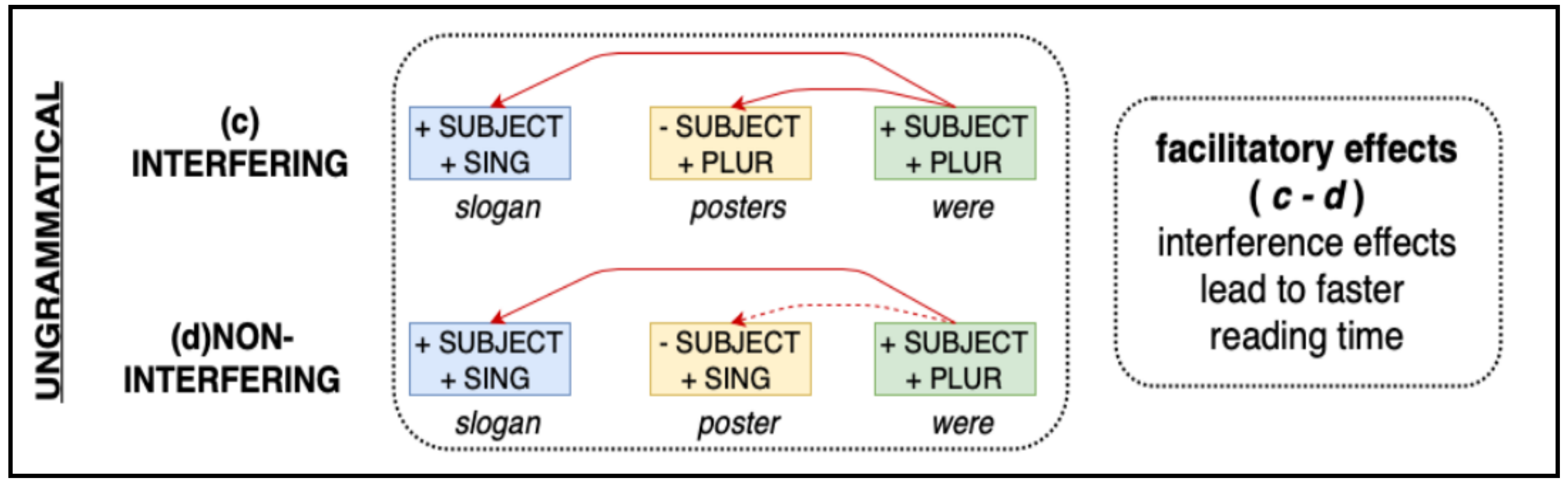
Current Proposal: The Full Picture

Computational Level

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



Psycholinguistic theories (?)



Algorithmic Level ONE

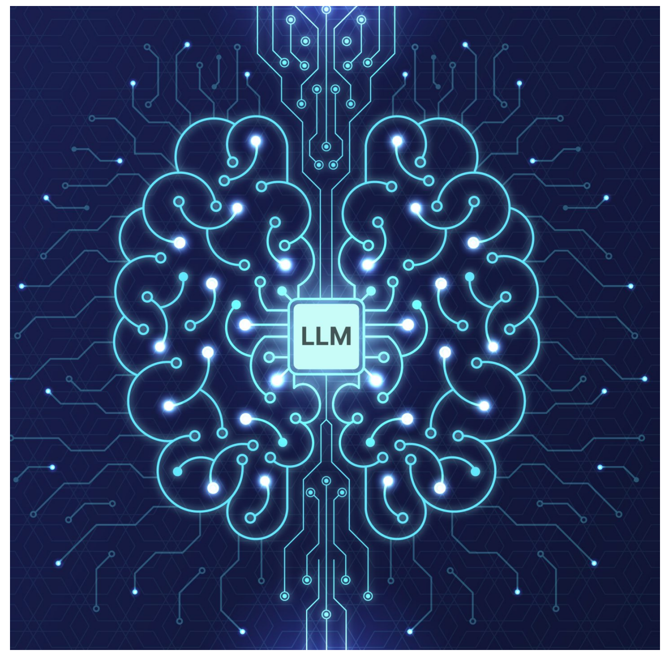
An "algorithm" of sentence processing

Algorithmic Level TWO

Current Proposal: The Full Picture

Computational Level

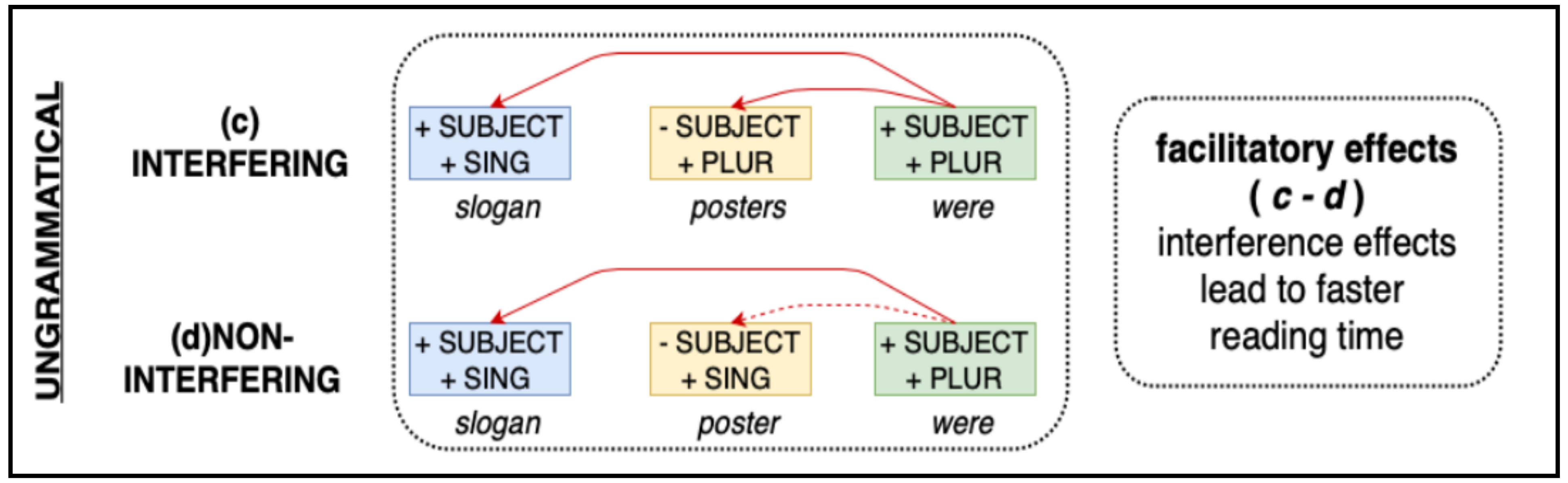
Language **in the broader design of intelligent systems**



Algorithmic Level ONE

Psycholinguistic theories (?)

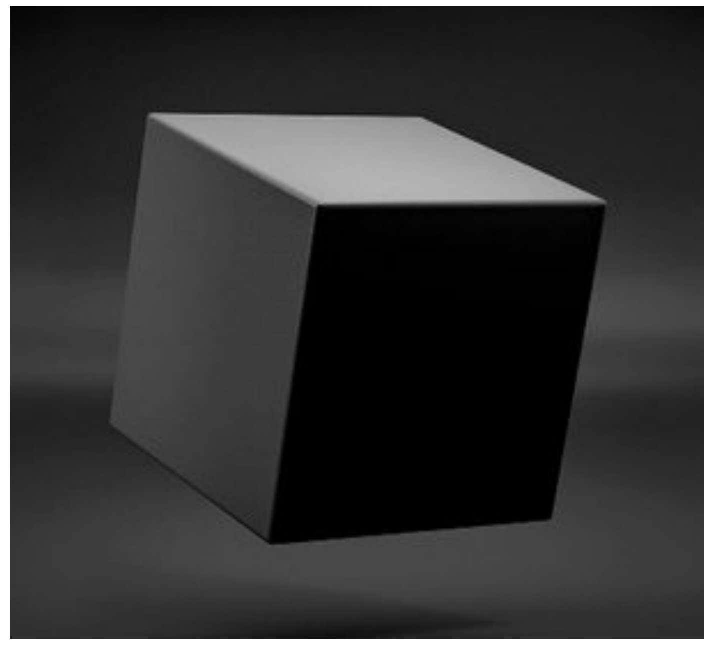
Functionally approximating

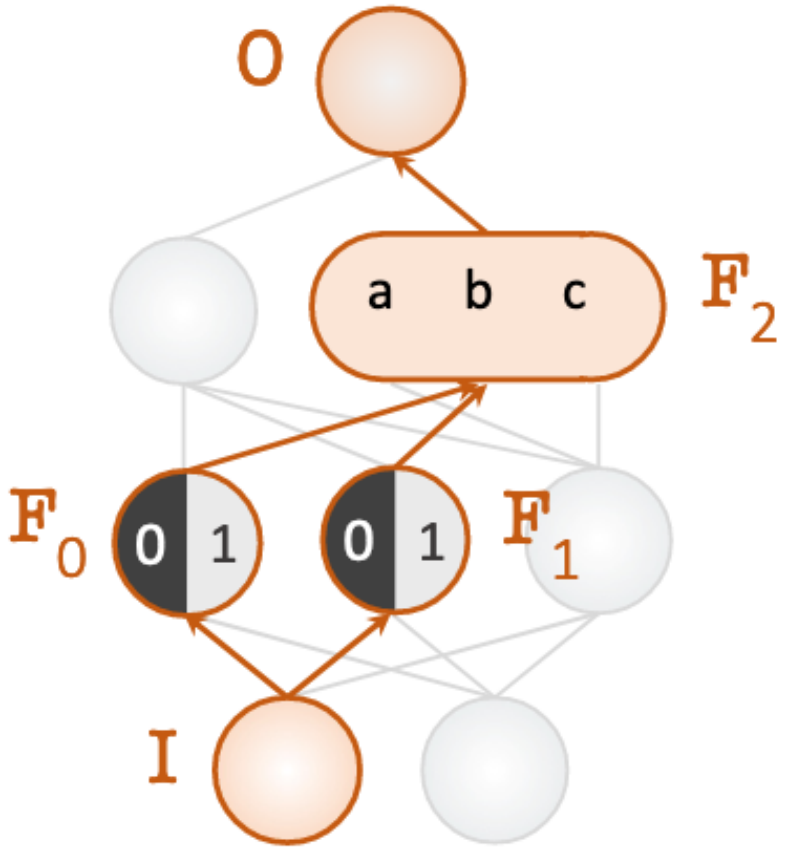


Functionally approximating

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!