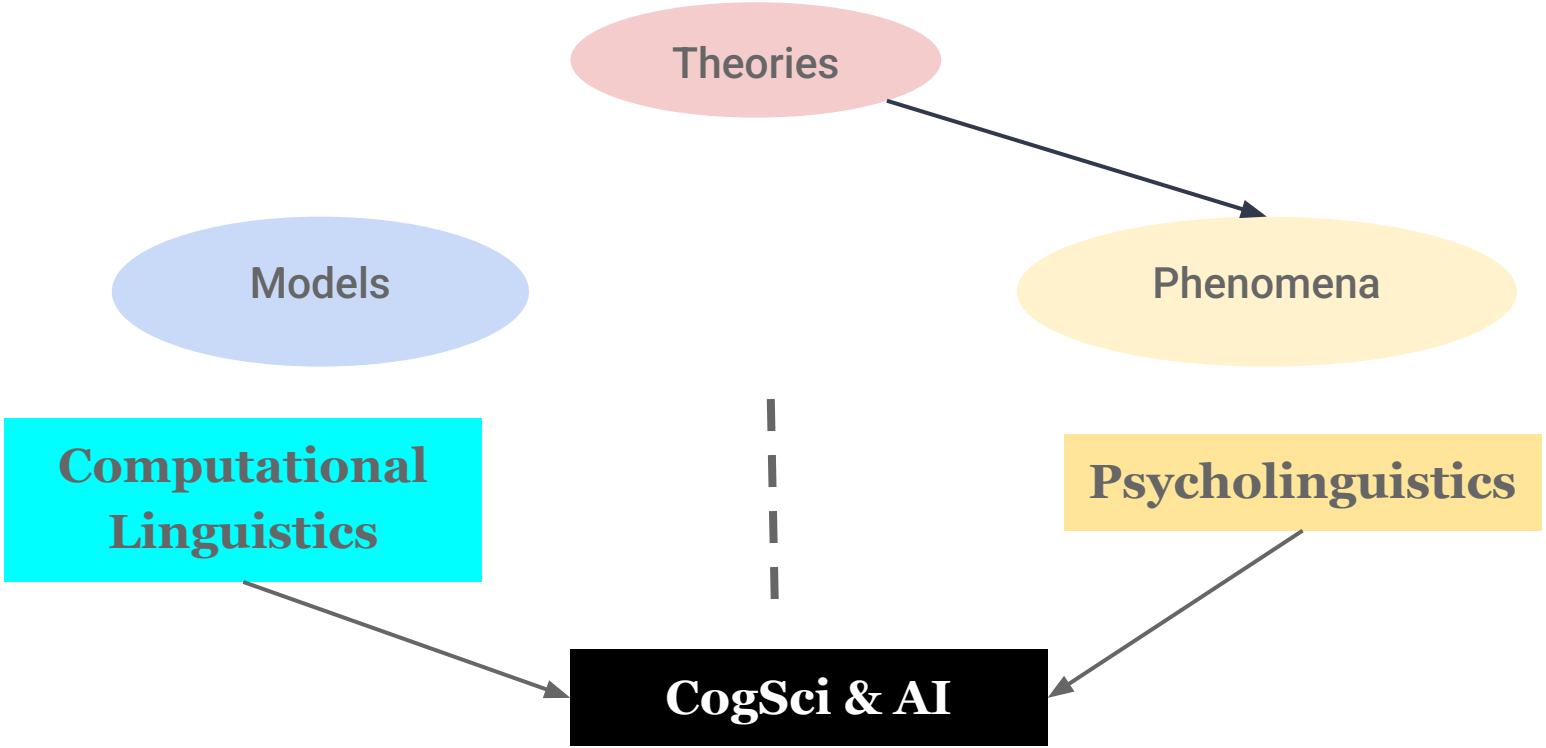
Is In-Context Learning a Type of Error-Driven Learning Mechanism?

Zhenghao “Herbert” Zhou
Yale University
Hollymartin Lab @ University of Edinburgh
December 4, 2024

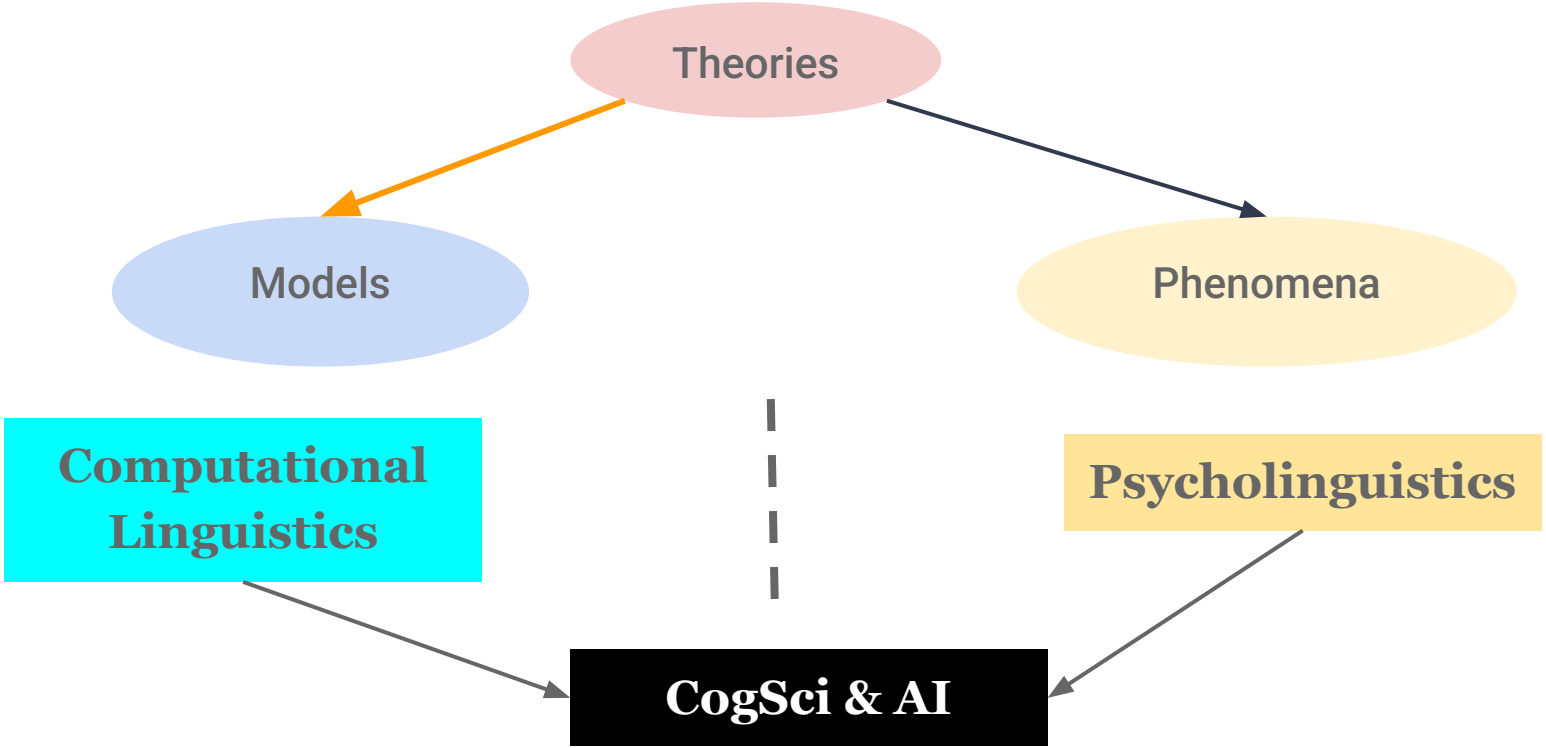
Evidence from the Inverse Frequency Effects in Structural Priming

Big Picture: Notes on Methodology

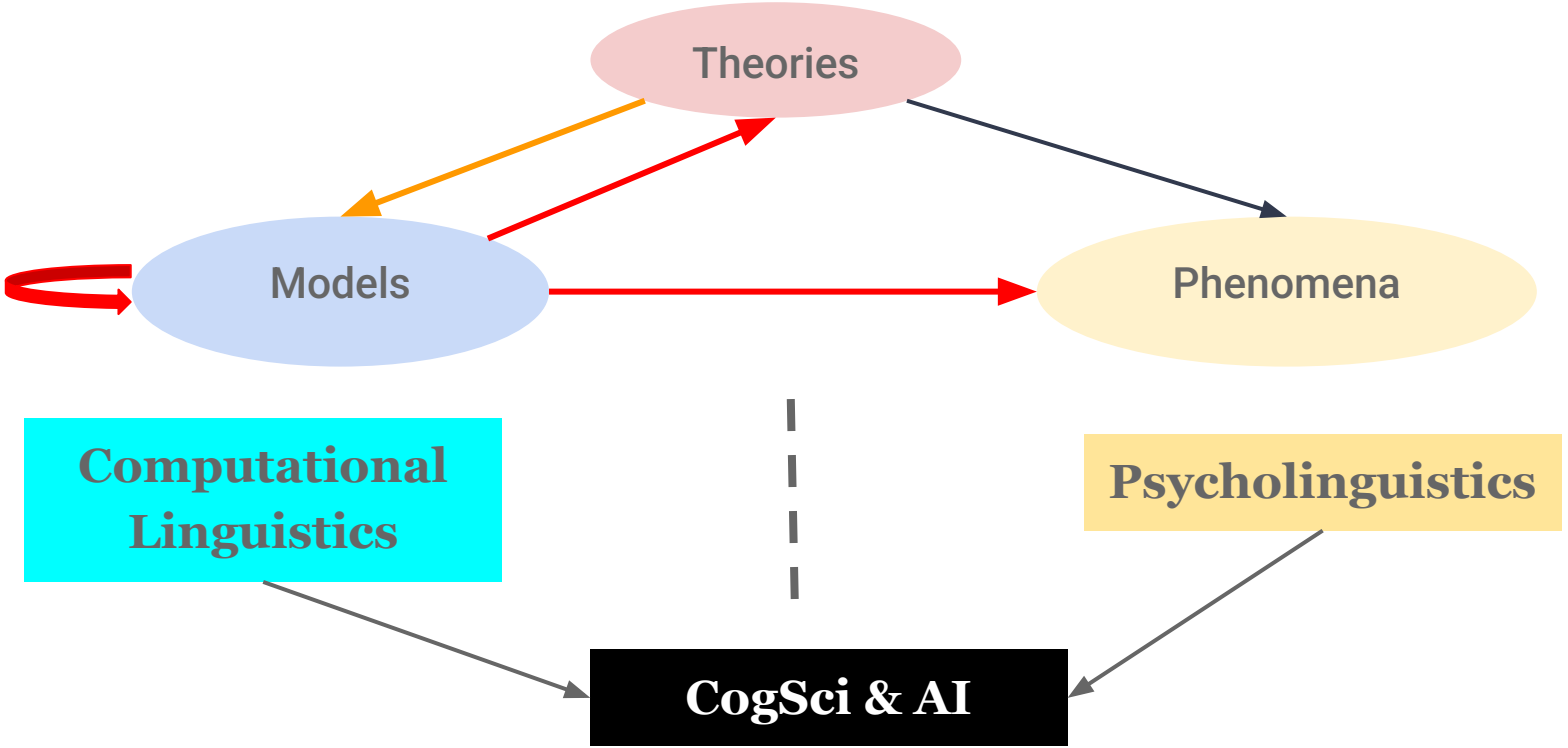
Big Picture: Notes on Methodology



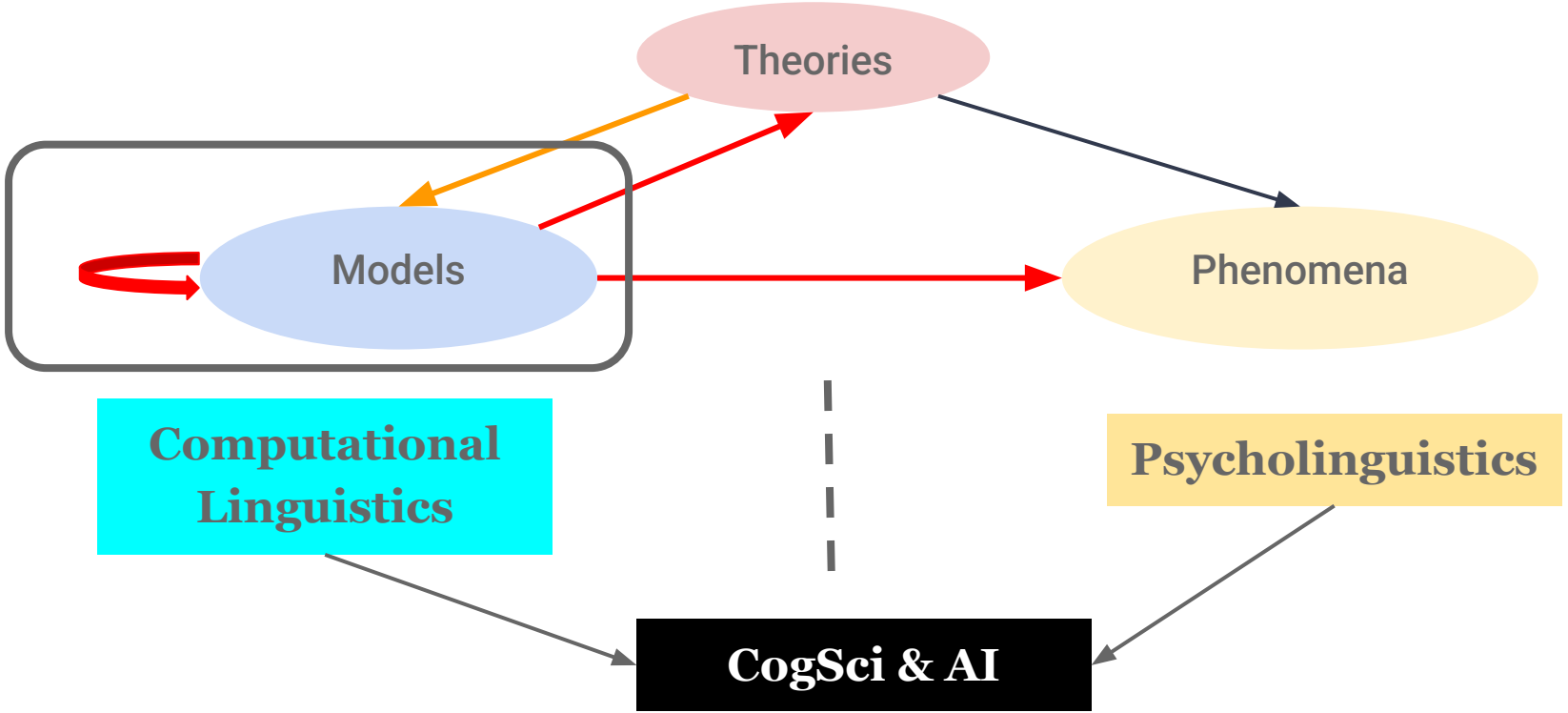
Big Picture: Notes on Methodology



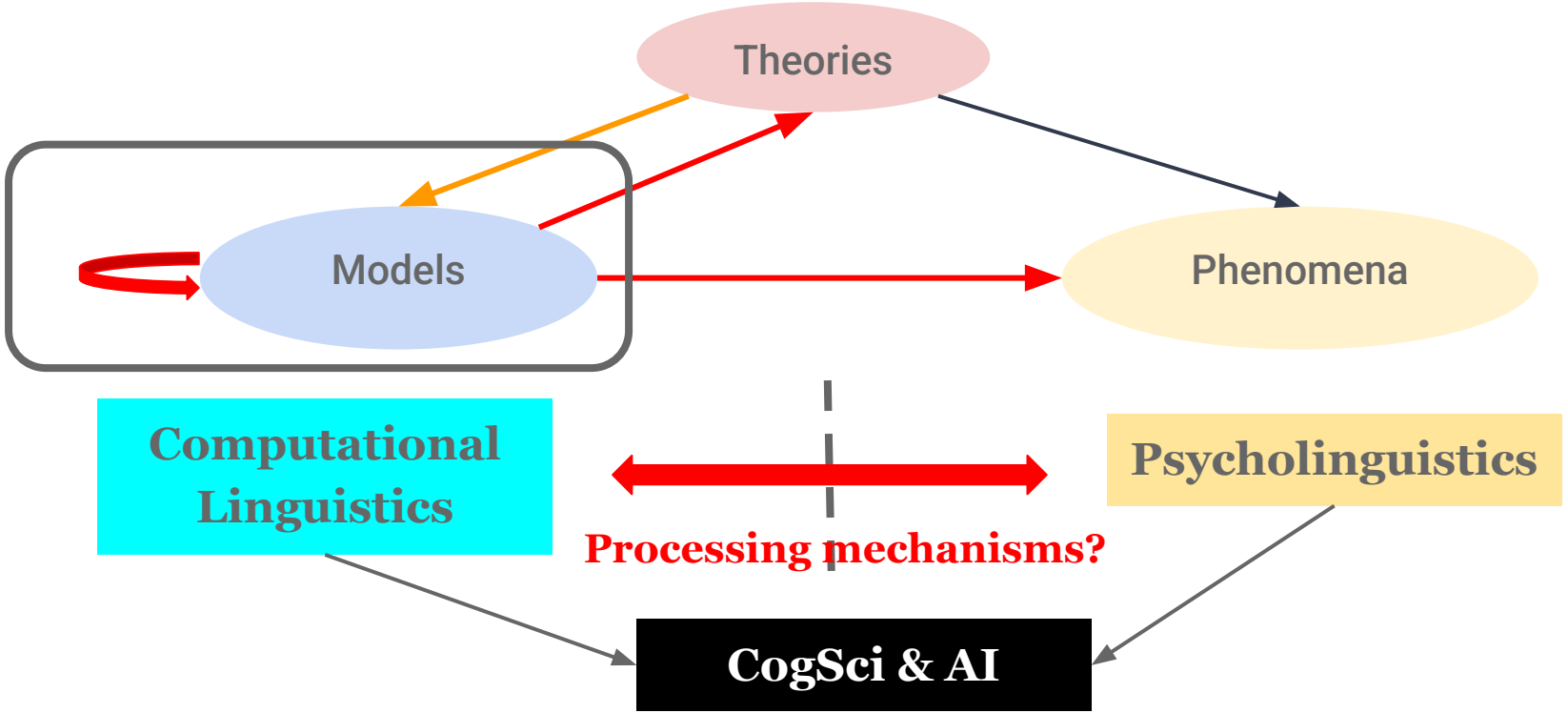
Big Picture: Notes on Methodology



Big Picture: Notes on Methodology



Big Picture: Notes on Methodology



Outline

In-context Learning (ICL) in LLMs

- *In-context* learning vs. *In-weights* learning
- ICL as a processing mechanism of LLMs;
- Is ICL functionally performing some error-driven learning?

Outline

In-context Learning (ICL) in LLMs

- *In-context* learning vs. *In-weights* learning
- ICL as a processing mechanism of LLMs;
- Is ICL functionally performing some error-driven learning?

Structural Priming

- Structural priming effect;
- The Inverse Frequency Effect (IFE);
- Two accounts: transient activation vs. implicit learning;

Outline

In-context Learning (ICL) in LLMs

- *In-context* learning vs. *In-weights* learning
- ICL as a processing mechanism of LLMs;
- Is ICL functionally performing some error-driven learning?

Structural Priming

- Structural priming effect;
- The Inverse Frequency Effect (IFE);
- Two accounts: transient activation vs. implicit learning;

Current Study

- Analogy between priming and ICL?
- Do LLMs show the IFE?
- IFE as a diagnostic of error-driven learning?

Outline

In-context Learning (ICL) in LLMs

- *In-context* learning vs. *In-weights* learning
- ICL as a processing mechanism of LLMs;
- Is ICL functionally performing some error-driven learning?

Structural Priming

- Structural priming effect;
- The Inverse Frequency Effect (IFE);
- Two accounts: transient activation vs. implicit learning;

Current Study

- Analogy between priming and ICL?
- Do LLMs show the IFE?
- IFE as a diagnostic of error-driven learning?

Discussion & Implications

- Larger models show stronger IFE;
- There is an *implicit* gradient component involved in ICL;
- Humans and LLMs share a similar processing mechanism;

Large Language Models (LLMs)

Model

Autoregressive Models

- GPT2 small: 117M
- GPT2 medium: 345M
- GPT2 large: 762M
- LLAMA2 7B: 7B
- LLAMA2 7B-chat: 7B
- LLAMA2 70B: 70B
- GPT3 davinci-002: 175B

https://jalammar.github.io/illustrated-gpt2/

Large Language Models (LLMs)

Model

Predicting the next word!

Autoregressive Models

- GPT2 small: 117M
- GPT2 medium: 345M
- GPT2 large: 762M
- LLAMA2 7B: 7B
- LLAMA2 7B-chat: 7B
- LLAMA2 70B: 70B
- GPT3 davinci-002: 175B

https://jalammar.github.io/illustrated-gpt2/

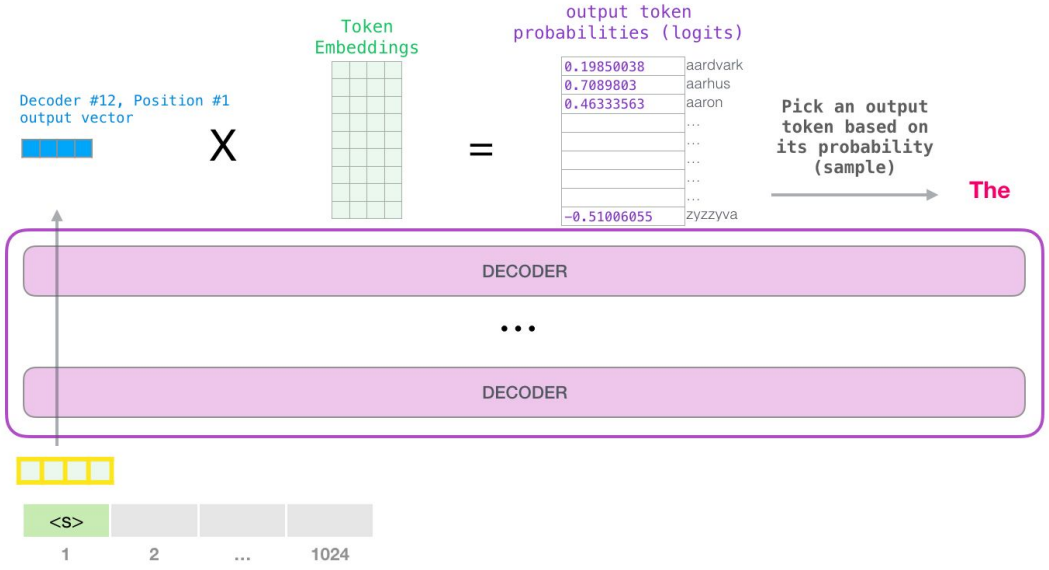
Large Language Models (LLMs)

Model

Predicting the next word!

Autoregressive Models

- GPT2 small: 117M
- GPT2 medium: 345M
- GPT2 large: 762M
- LLAMA2 7B: 7B
- LLAMA2 70B: 70B
- GPT3 davinci-002: 175B



https://jalammar.github.io/illustrated-gpt2/

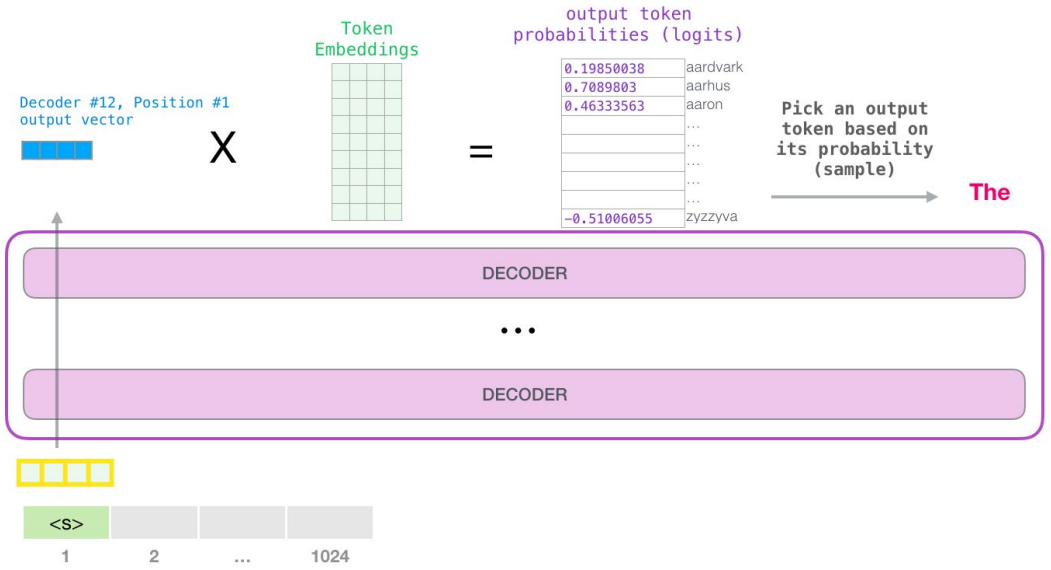
Large Language Models (LLMs)

Model

Predicting the next word!

Autoregressive Models

- GPT2 small: 117M
- GPT2 medium: 345M
- GPT2 large: 762M
- LLAMA2 7B: 7B
- LLAMA2 70B: 70B
- GPT3 davinci-002: 175B



Parameters: weights and biases; updated during training; ⇒ long-term memory~ish;
Activations: temporarily variables generated and modified during generation (using LMs);

https://jalammar.github.io/illustrated-gpt2/

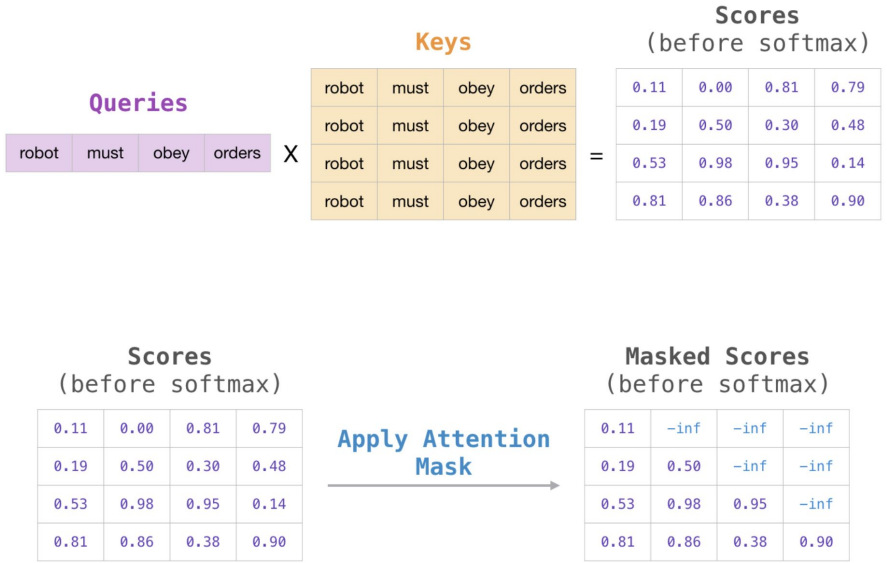
Large Language Models (LLMs)

Model

Predicting the next word!

Autoregressive Models

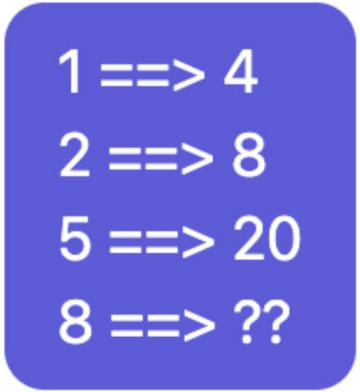
- GPT2 small: 117M
- GPT2 medium: 345M
- GPT2 large: 762M
- LLAMA2 7B: 7B
- LLAMA2 7B-chat: 7B
- LLAMA2 70B: 70B
- GPT3 davinci-002: 175B



Parameters: weights and biases; updated during training; ⇒ long-term memory~ish;
Activations: temporarily variables generated and modified during generation (using LMs);

https://jalammar.github.io/illustrated-gpt2/

In-context Learning in LLMs



Model

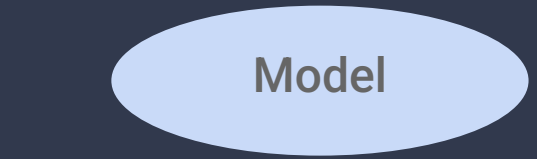
[Brown et al. 2020]

In-context Learning in LLMs

&() --> ()&
234 --> 342
@#$ --> #$@
rte--->

[Brown et al. 2020]

In-context Learning in LLMs



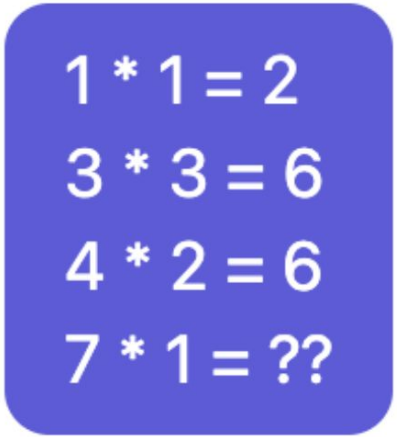
1 * 1 = 2
3 * 3 = 6
4 * 2 = 6
7 * 1 = ??

[Brown et al. 2020]

In-context Learning in LLMs

Model

In-context Learning: having a demonstration (i.e. several <example, answer> pairs) of a (implicitly defined) task increases the model performance.

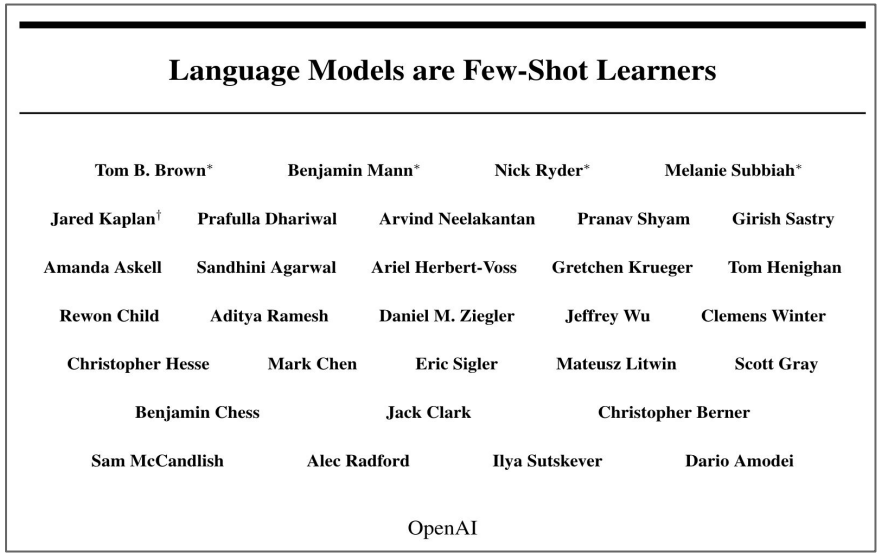


[Brown et al. 2020]

In-context Learning in LLMs

Model

In-context Learning: having a demonstration (i.e. several <example, answer> pairs) of a (implicitly defined) task increases the model performance.



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

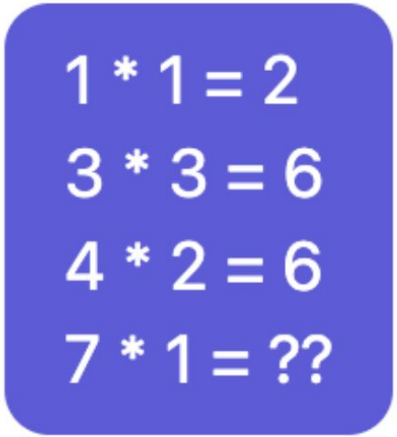
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

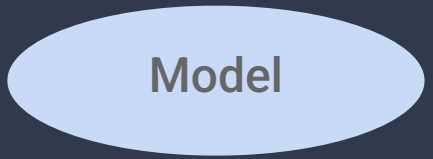
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI



[Brown et al. 2020]

ICL vs. In-weights Learning



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
```

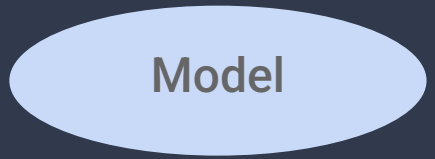
Few-shot

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]

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.

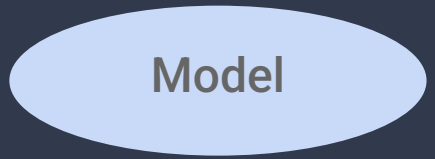
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ←
4  plush girafe => girafe peluche      ←
5  cheese =>                           ← prompt
```

[Brown et al. 2020]

In-*context* Learning:

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

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:

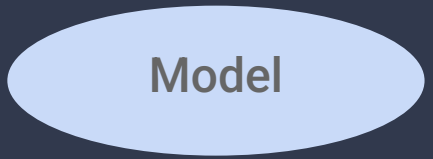
- No gradient updates;
- Rapid: from a few examples;
- One-shot / Few-shot learning;

Fine-tuning

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

```
1  sea otter => loutre de mer          ← example #1
        ↓
   gradient update
        ↓
1  peppermint => menthe poivrée        ← example #2
        ↓
   gradient update
        ↓
        ...
        ↓
1  plush giraffe => girafe peluche     ← example #N
        ↓
   gradient update
        ↓
1  cheese =>                           ← prompt
```

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):

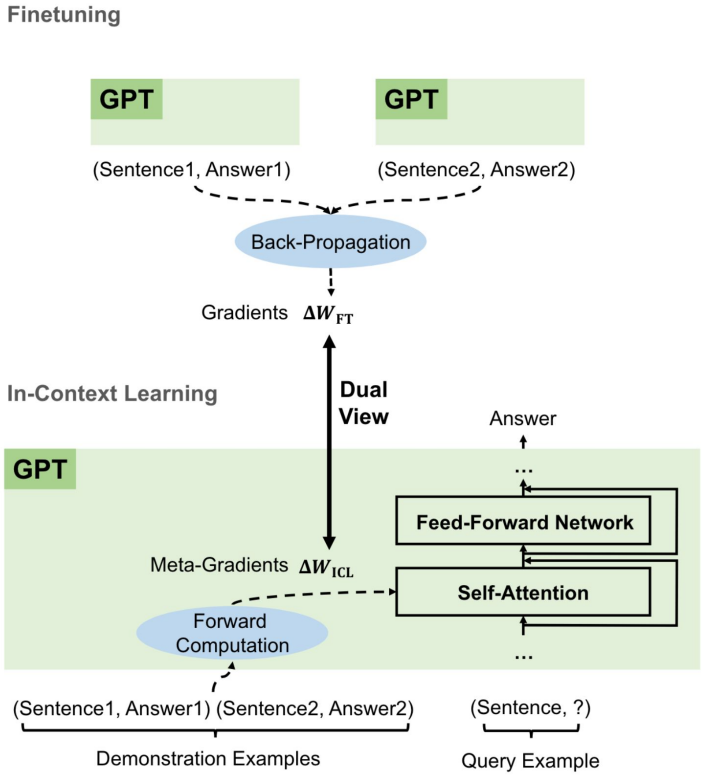
- Gradient-based;
- Slow: need many examples;
- Standard supervised learning;

Fine-tuning

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

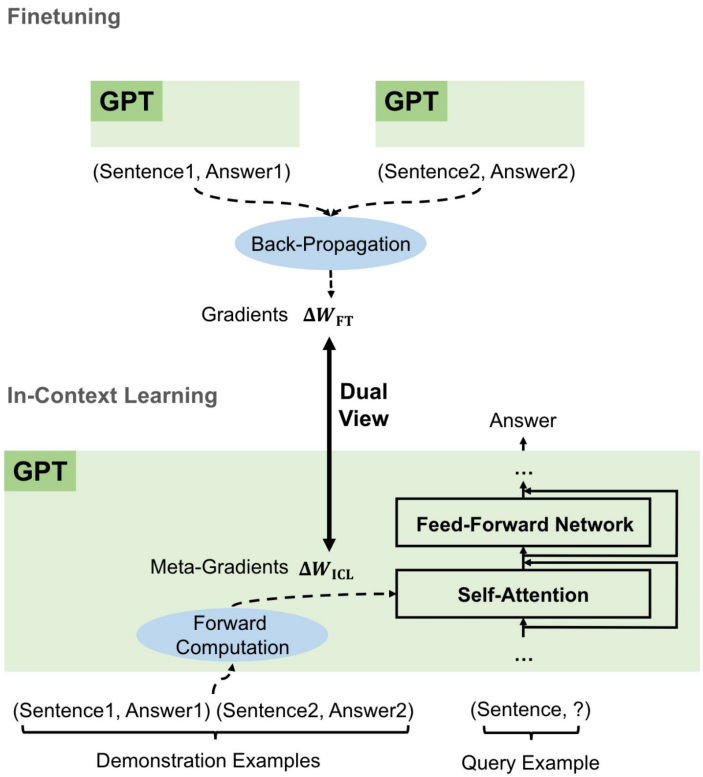
```
1  sea otter => loutre de mer          ← example #1
        ↓
   gradient update
        ↓
1  peppermint => menthe poivrée        ← example #2
        ↓
   gradient update
        ↓
   • • •
        ↓
1  plush giraffe => girafe peluche     ← example #N
        ↓
   gradient update
        ↓
1  cheese =>                           ← prompt
```

Open Question: ICL ≈(functional) Gradient Descent?



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

Open Question: ICL ≈(functional) Gradient Descent?

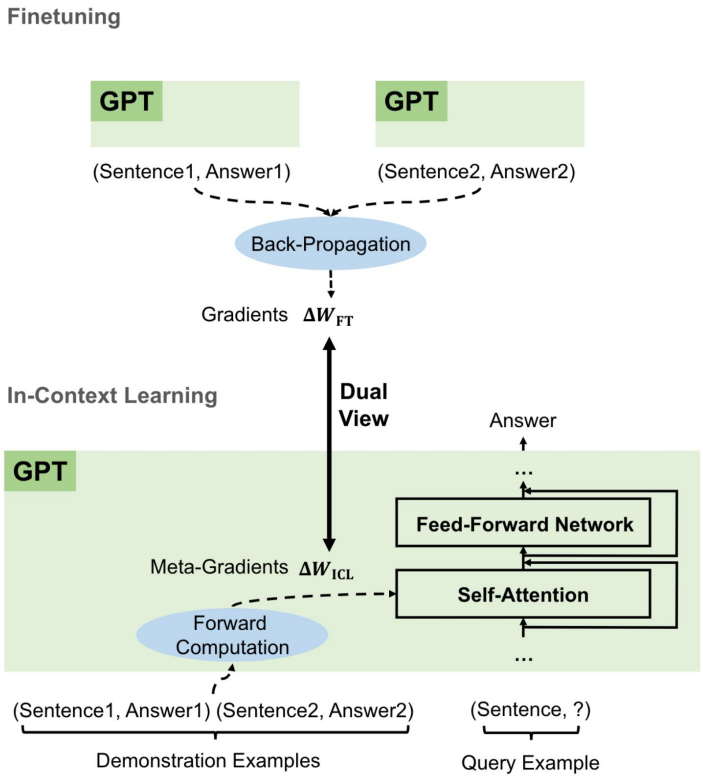


In-context Learning as implicitly performing gradient descent? — *in principle*, yes…

- **ICL performs implicit Bayesian inference;**
- **ICL functionally performs gradient descent;**
- **ICL as a meta-optimization process equivalent to implicit fine-tuning;**

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

Open Question: ICL ≈(functional) Gradient Descent?



In-context Learning as implicitly performing gradient descent? — *in principle*, yes…

- **ICL performs implicit Bayesian inference;**
- **ICL functionally performs gradient descent;**
- **ICL as a meta-optimization process equivalent to implicit fine-tuning;**

Current Case Study: Is there an error-based learning process in the forward pass? — testing with off-the-shelf LLMs and natural language!

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

Interim Summary 1

- *In-context* learning vs. *In-weights* learning
- ICL as a processing mechanism of LLMs;
- Is ICL functionally performing some error-driven learning?

Structural Priming

Phenomenon

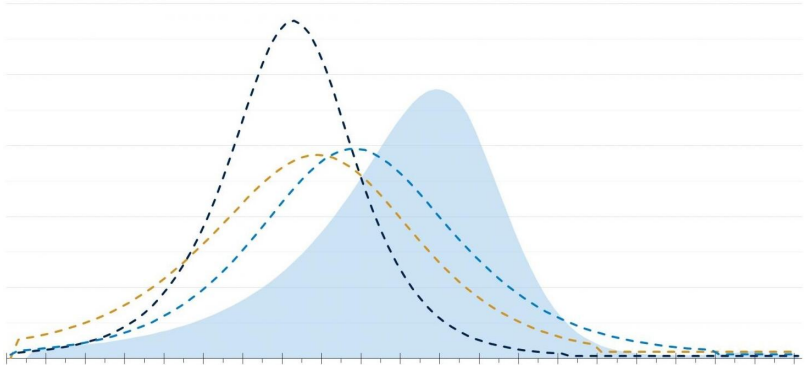
Linguistic Adaptation: the linguistic knowledge representations that are used for language processing **change** in response to language input.

E.g. [Bock 1986, Chang 2012]

Structural Priming

Phenomenon

Linguistic Adaptation: the linguistic knowledge representations that are used for language processing **change** in response to language input.



E.g. [Bock 1986, Chang 2012]

Structural Priming

Phenomenon

Linguistic Adaptation: the linguistic knowledge representations that are used for language processing **change** in response to language input.

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

E.g. [Bock 1986, Chang 2012]

Structural Priming

Phenomenon

Linguistic Adaptation: the linguistic knowledge representations that are used for language processing **change** in response to language input.

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

Our focus: Double Object (DO) vs. Prepositional Dative (PD) for ditransitive predicates.

E.g. [Bock 1986, Chang 2012]

Structural Priming

Phenomenon

Linguistic Adaptation: the linguistic knowledge representations that are used for language processing **change** in response to language input.

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

Our focus: Double Object (DO) vs. Prepositional Dative (PD) for ditransitive predicates.

- DO: *Alice sent Bob a letter*.
- PD: *Alice sent a letter to Bob*.

E.g. [Bock 1986, Chang 2012]

Inverse Frequency Effect (IFE)

Phenomenon

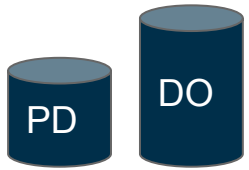
Inverse Frequency Effect: the less preferred (lower frequency) syntactic structure causes a stronger priming effect than the more preferred (higher frequency) structural alternative.

E.g. [Jaeger & Snider 2007]

Inverse Frequency Effect (IFE)

Phenomenon

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



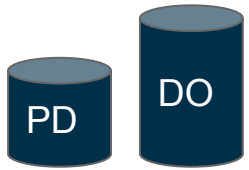
Verb Bias:
buy is biased towards DO
design towards PD

E.g. [Jaeger & Snider 2007]

Inverse Frequency Effect (IFE)

Phenomenon

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 towards PD

$T_{\text{PD}} | P_{\text{DO}}$

Prime in DO Structure **Target in PD Structure**

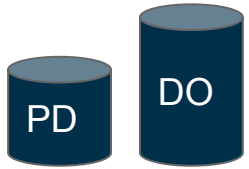
A doctor bought a chief a plate. The secretary drew the card for the band.
A doctor designed a chief a plate. The secretary drew the card for the band.

E.g. [Jaeger & Snider 2007]

Inverse Frequency Effect (IFE)

Phenomenon

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 towards PD

$T_{\text{PD}} | P_{\text{DO}}$

Prime in DO Structure **Target in PD Structure**

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

Greater priming effect!

E.g. [Jaeger & Snider 2007]

Inverse Frequency Effect (IFE)

Phenomenon

$$T_{\text{PD}} \mid P_{\text{DO}}$$

Unprimed log probability = -50

Prime in DO Structure → **Target in PD Structure**

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

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

Inverse Frequency Effect (IFE)

Phenomenon

$$T_{PD} \mid P_{DO}$$

Unprimed log probability = -50

Prime in DO Structure — **Target in PD Structure**

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

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

Prime Verb

Bring
Buy
Find
Draw
Design

Inverse Frequency Effect (IFE)

Phenomenon

$$T_{PD} | P_{DO}$$

Unprimed log probability = -50

Prime in DO Structure — **Target in PD Structure**

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

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

Prime Verb	Verb PD Bias	Primed log probability (priming magnitude)
Bring	*0.23*	*-50.5 (0.5)*
Buy	*0.27*	*-51.1 (1.1)*
Find	*0.41*	*-51.8 (1.8)*
Draw	*0.52*	*-52.2 (2.2)*
Design	*0.77*	*-52.9 (2.9)*

Inverse Frequency Effect (IFE)

Phenomenon

$$T_{PD} | P_{DO}$$

Unprimed log probability = -50

Prime in DO Structure → **Target in PD Structure**

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

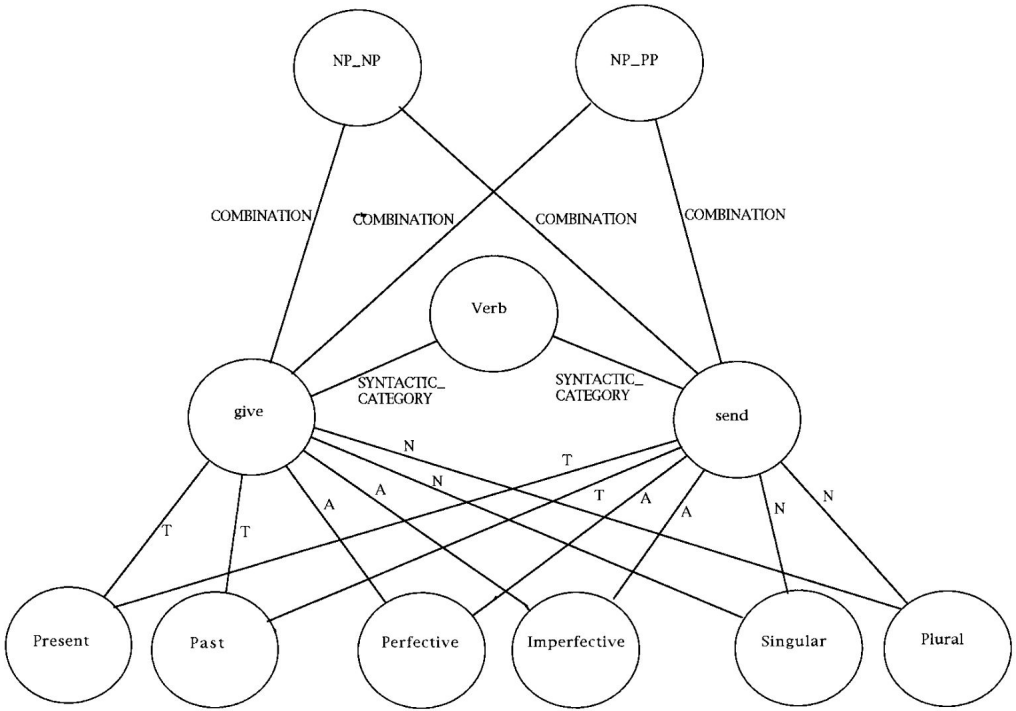
A doctor designed a chief a plate. The secretary drew the card for the band.

Prime Verb	Verb PD Bias	Primed log probability (priming magnitude)
Bring	*0.23*	*-50.5 (0.5)*
Buy	*0.27*	*-51.1 (1.1)*
Find	*0.41*	*-51.8 (1.8)*
Draw	*0.52*	*-52.2 (2.2)*
Design	*0.77*	*-52.9 (2.9)*

When priming in DO structure → *larger PD biases* → *a greater priming effect*

Transient Activation

Theory 1

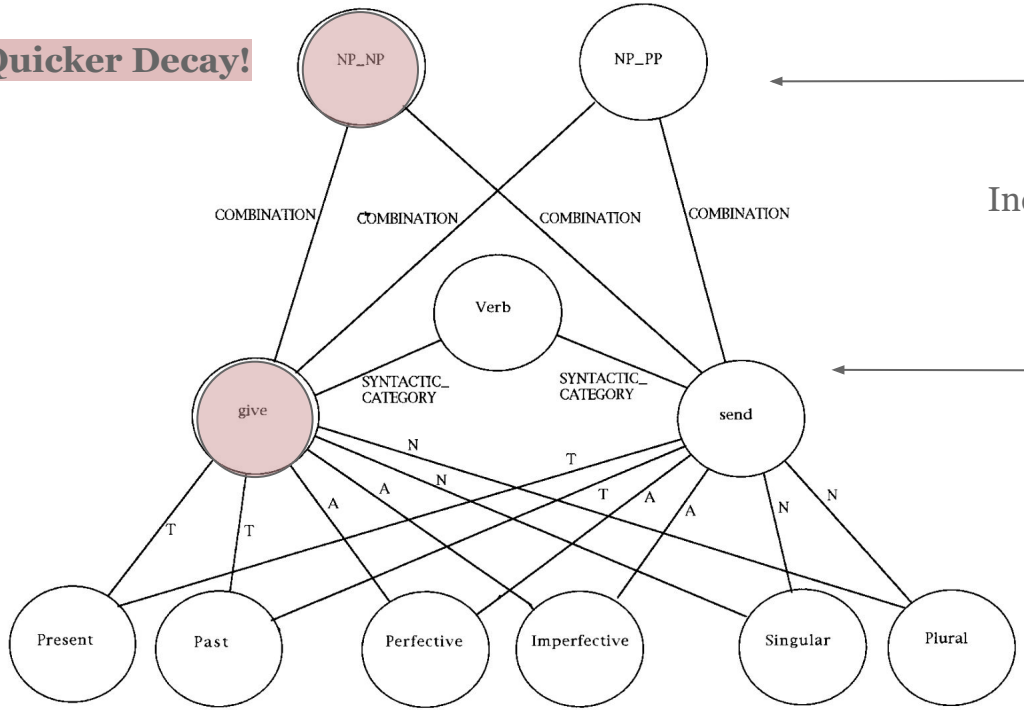


[Pickering & Branigan 1998]

Transient Activation

Theory 1

Quicker Decay!



Structural Nodes

Increasing the activation values of the nodes.

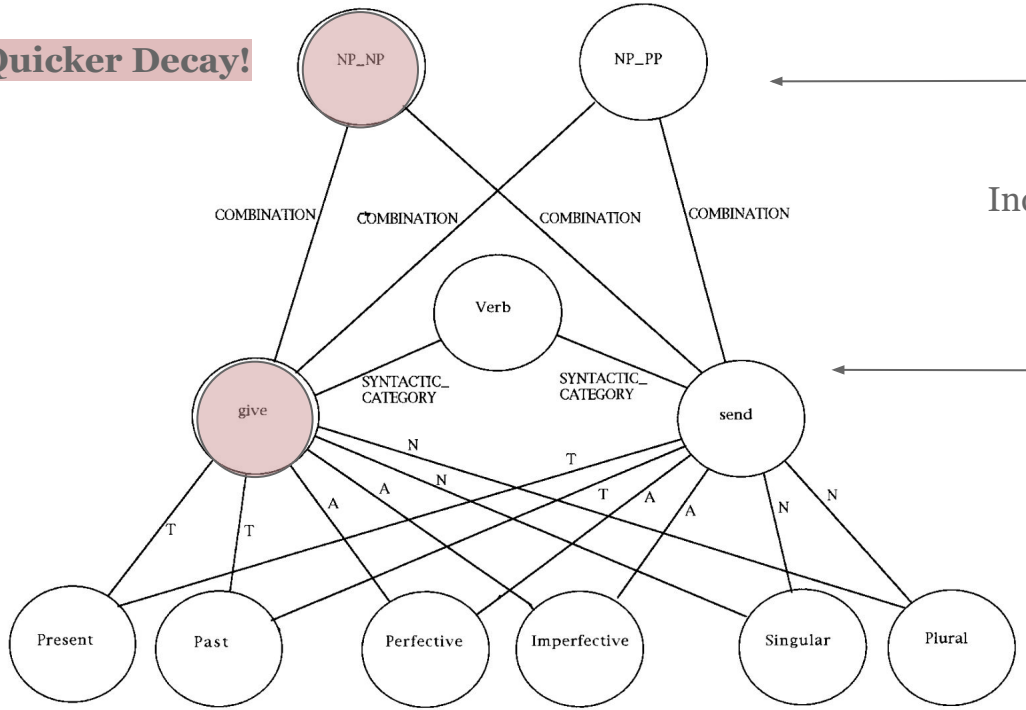
Lexical Nodes

[Pickering & Branigan 1998]

Transient Activation

Theory 1

Quicker Decay!



Structural Nodes

Increasing the activation values of the nodes.

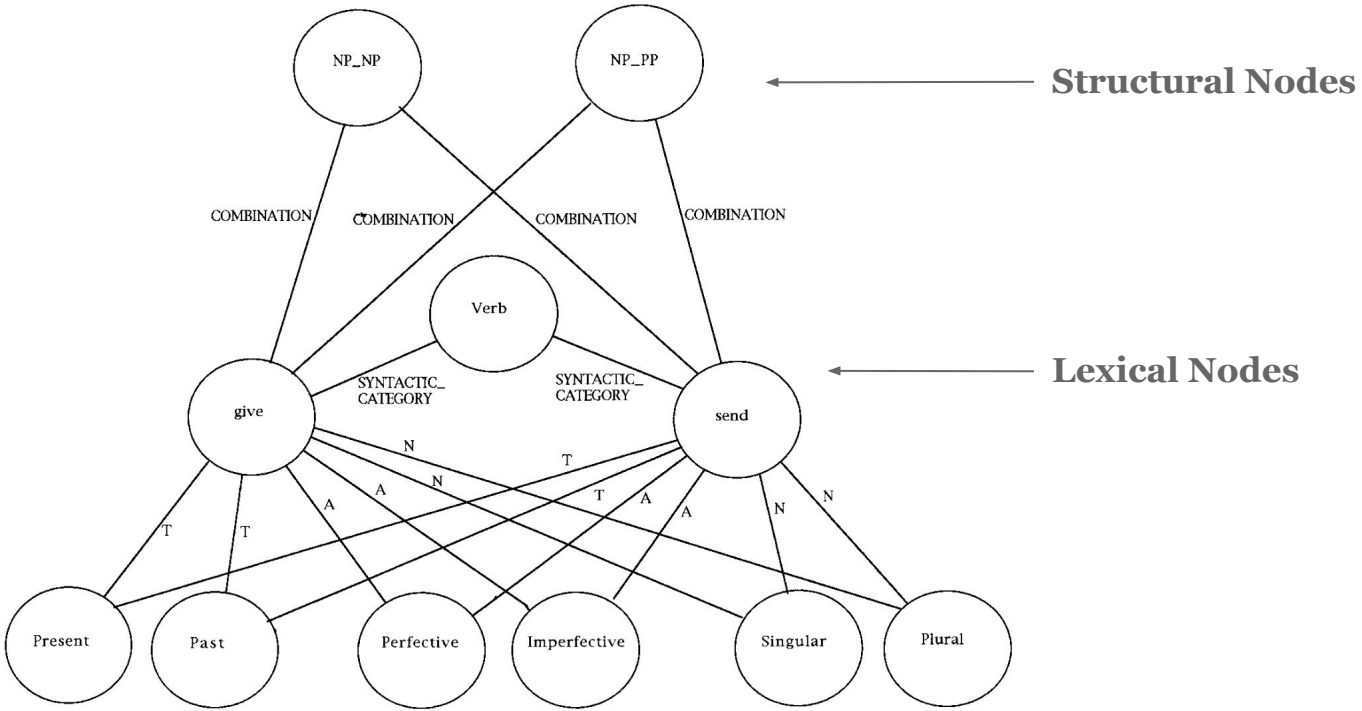
Lexical Nodes

- **Activation instead of learning**
- **Explaining short-term phenomena**
- **Inverse Frequency:** ❌

[Pickering & Branigan 1998]

Implicit Learning

Theory 2

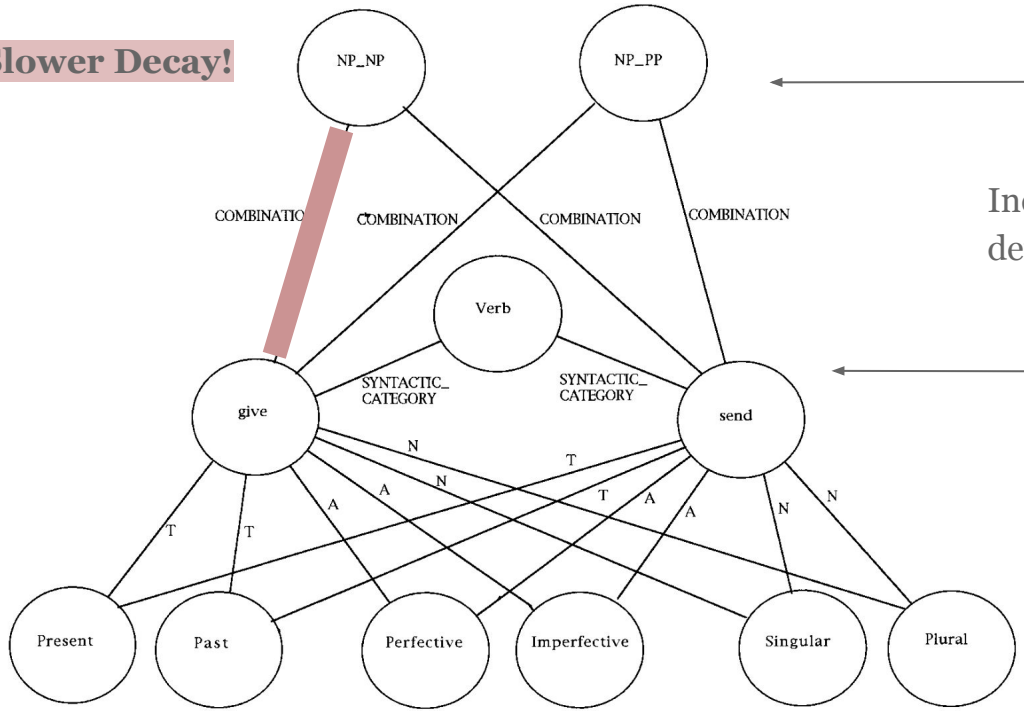


E.g. [Chang et al. 2006]

Implicit Learning

Theory 2

Slower Decay!



Structural Nodes

Increasing link weights *proportional to* the detected error from verb biases.

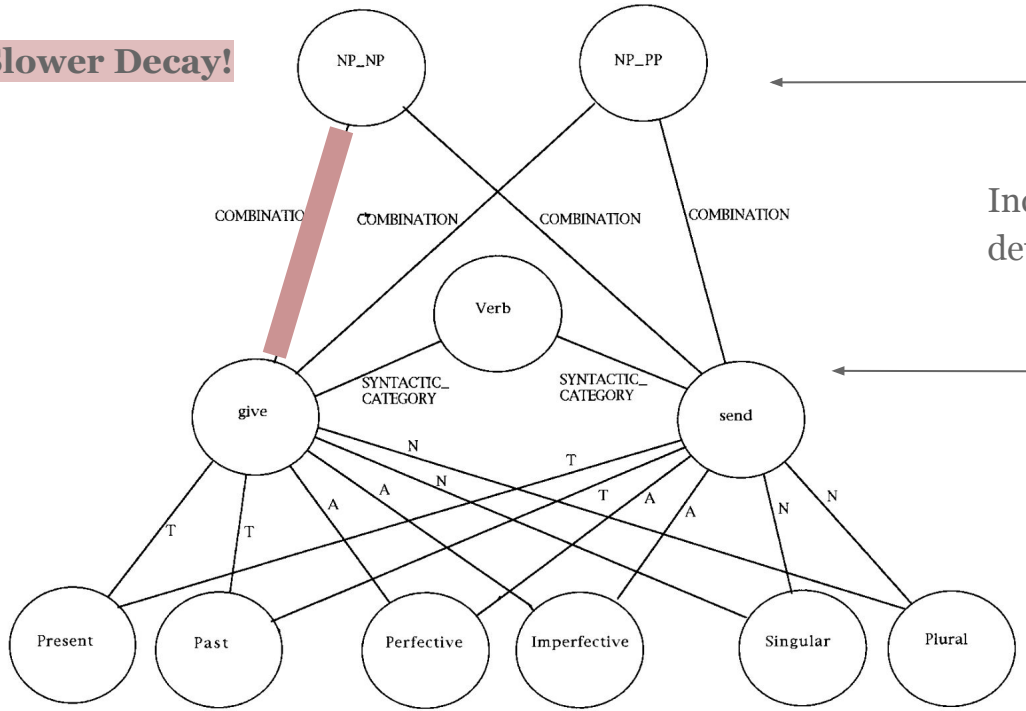
Lexical Nodes

E.g. [Chang et al. 2006]

Implicit Learning

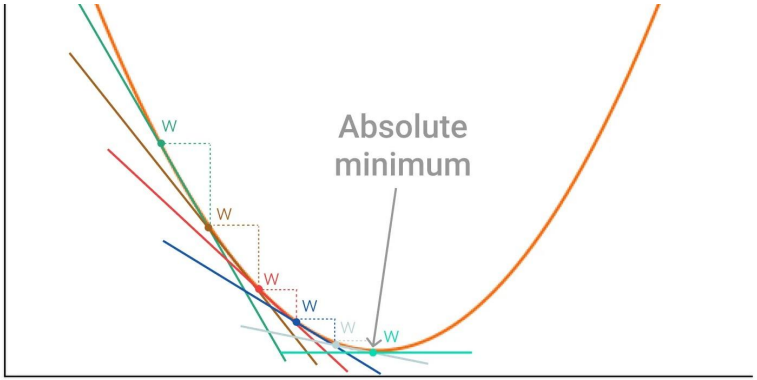
Theory 2

Slower Decay!



Structural Nodes

Increasing link weights *proportional to* the detected error from verb biases.

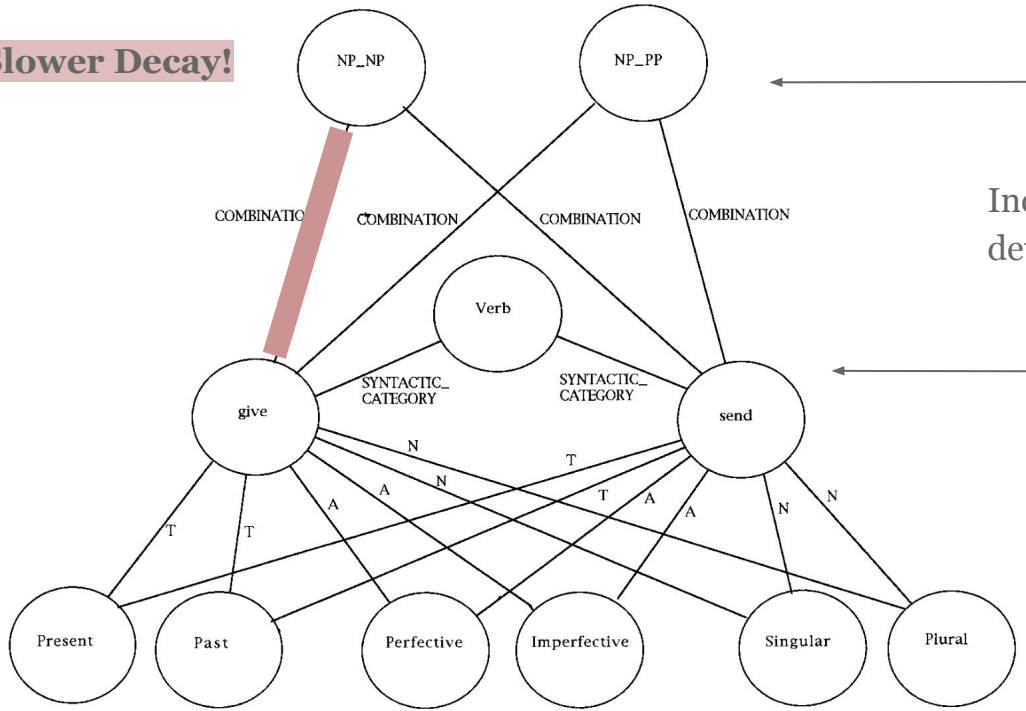


E.g. [Chang et al. 2006]

Implicit Learning

Theory 2

Slower Decay!



Structural Nodes

Increasing link weights *proportional to* the detected error from verb biases.

Lexical Nodes

- **Error-driven gradient update**
- **Explaining long-term Phenomena**
- **Inverse Frequency:** ✅

E.g. [Chang et al. 2006]

Interim Summary 2

In-context Learning (ICL) in LLMs | Structural Priming | Current Study | Discussion & Implications

- *In-context* learning vs. *In-weights* learning
- ICL as a processing mechanism of LLMs;
- Is ICL functionally performing some error-driven learning?

- Structural priming effect;
- The Inverse Frequency Effect (IFE);
- Two accounts: transient activation vs. implicit learning;

ICL as Structural Priming?

ICL as Structural Priming?

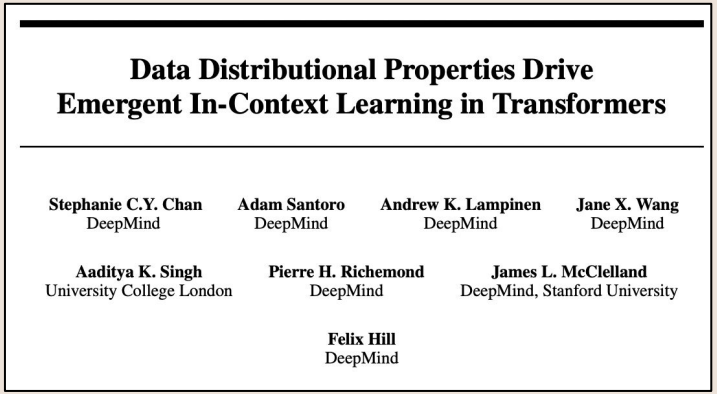
- Instead of viewing "A terrible movie. → negative" as an input-output pair, we can view it as a single "sentence" with a particular *structure*:
- Structure = movie review + arrow + sentiment label

A terrible movie. → *negative*

- Framed this way, in-context learning is structural priming!

LMs are shaped by what they are trained on!

LMs are shaped by what they are trained on!

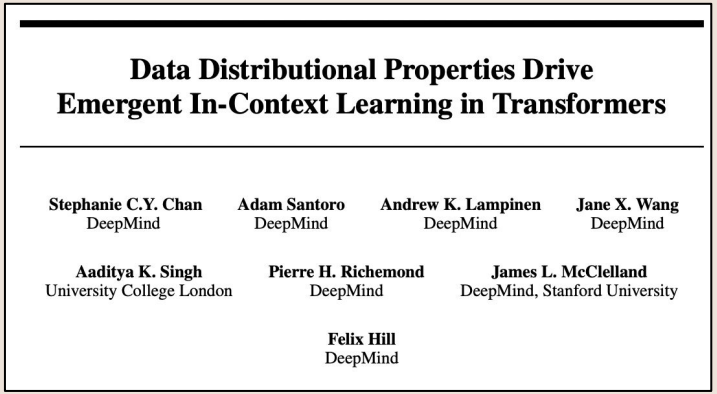


Data Distributional Properties Drive Emergent In-Context Learning in Transformers

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!

LMs are shaped by what they are trained on!

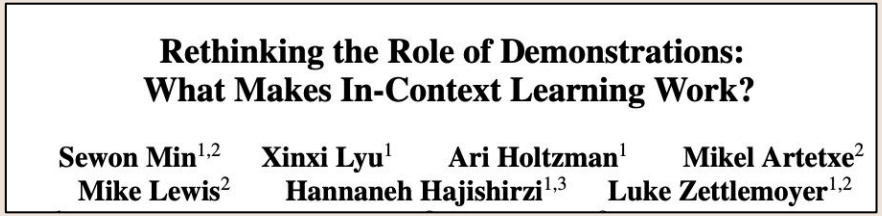


Data Distributional Properties Drive Emergent In-Context Learning in Transformers

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]

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?

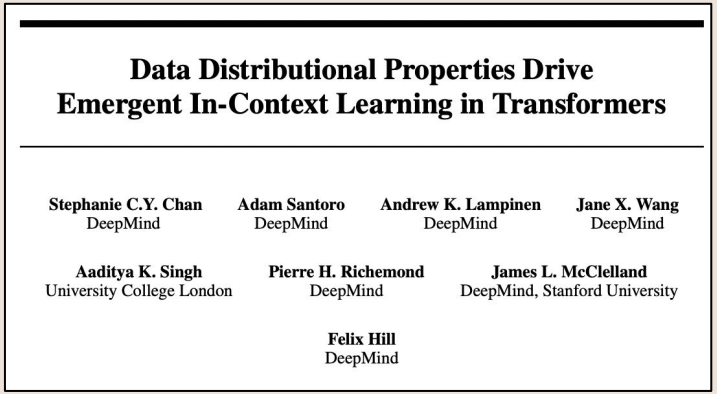
Sewon Min[1,2] Xinxi Lyu[1] Ari Holtzman[1] Mikel Artetxe[2]
Mike Lewis[2] Hannaneh Hajishirzi[1,3] Luke Zettlemoyer[1,2]

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]

LMs are shaped by what they are trained on!

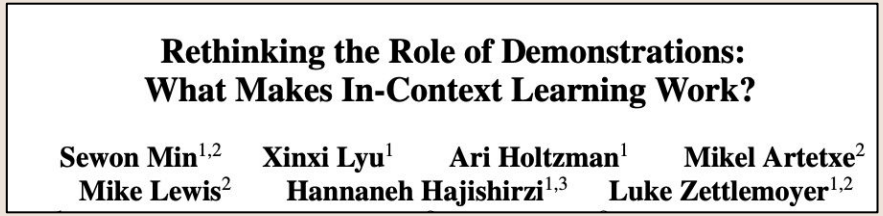


Data Distributional Properties Drive Emergent In-Context Learning in Transformers

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!

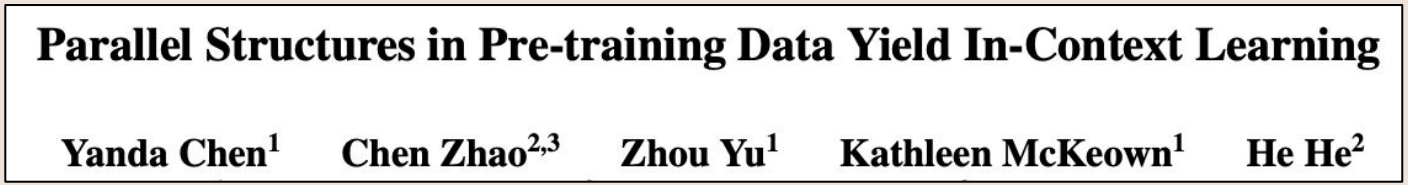
Compositional Structures!



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!

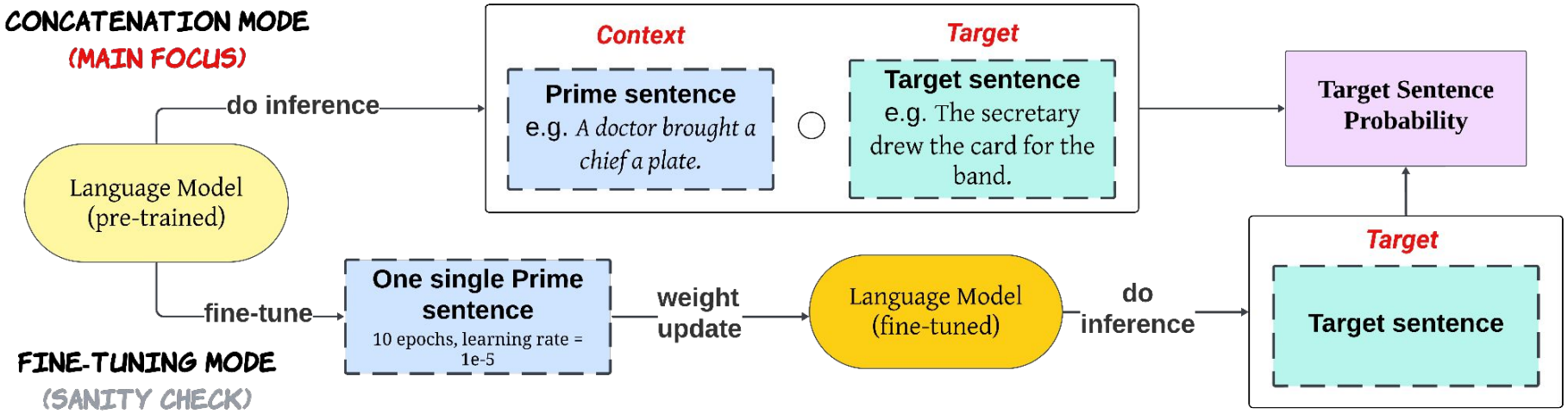


Parallel Structures in Pre-training Data Yield In-Context Learning

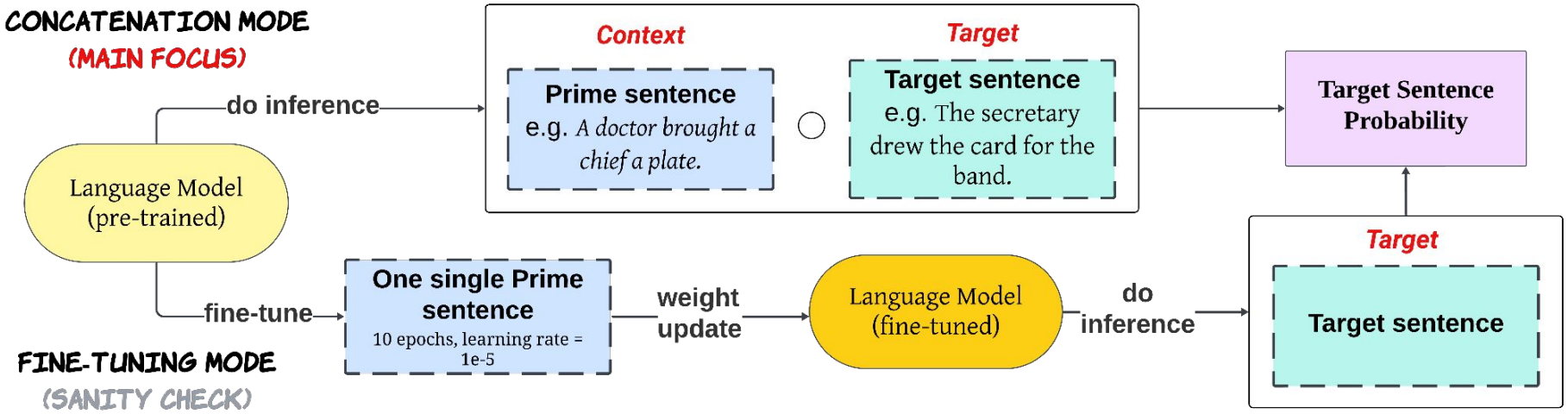
Yanda Chen[1] Chen Zhao[2,3] Zhou Yu[1] Kathleen McKeown[1] He He[2]

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

Current Approach: Overview



Current Approach: Overview



Assumption from Priming Theories: only some error-driven learning mechanism could lead to the IFE.

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

Corpus

Simulations

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

Corpus

Simulations

- 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

Simulations

- 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.**

$$t_{\mathrm{PD}} | p_{\mathrm{PD}}, t_{\mathrm{PD}} | p_{\mathrm{DO}}, t_{\mathrm{DO}} | p_{\mathrm{PD}}, t_{\mathrm{DO}} | p_{\mathrm{DO}}$$

Prime: *A doctor brought a plate to a chief.* **Target**: *The secretary drew the card for the band.*

Quantifying Verb Biases and the IFE

Simulations

Verb Bias of verb *V* on structure *X*:

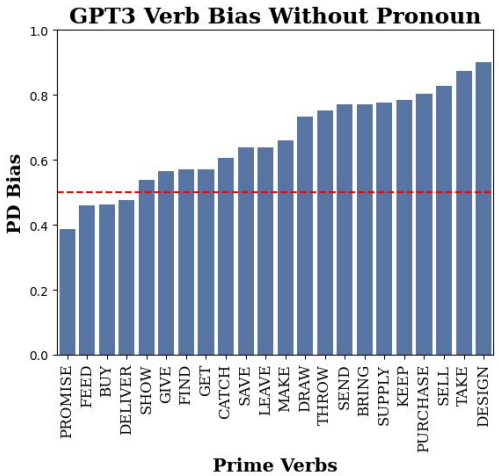
$$bias(V, \text{PD}) = \frac{1}{|\mathcal{S}_V|} \sum_{t_{\text{PD}} \in \mathcal{S}_V} \frac{\mathcal{P}(t_{\text{PD}})}{\mathcal{P}(t_{\text{PD}}) + \mathcal{P}(t_{\text{DO}})}$$

Quantifying Verb Biases and the IFE

Simulations

Verb Bias of verb *V* on structure *X*:

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



Quantifying Verb Biases and the IFE

Simulations

Verb Bias of verb *V* on structure *X*:

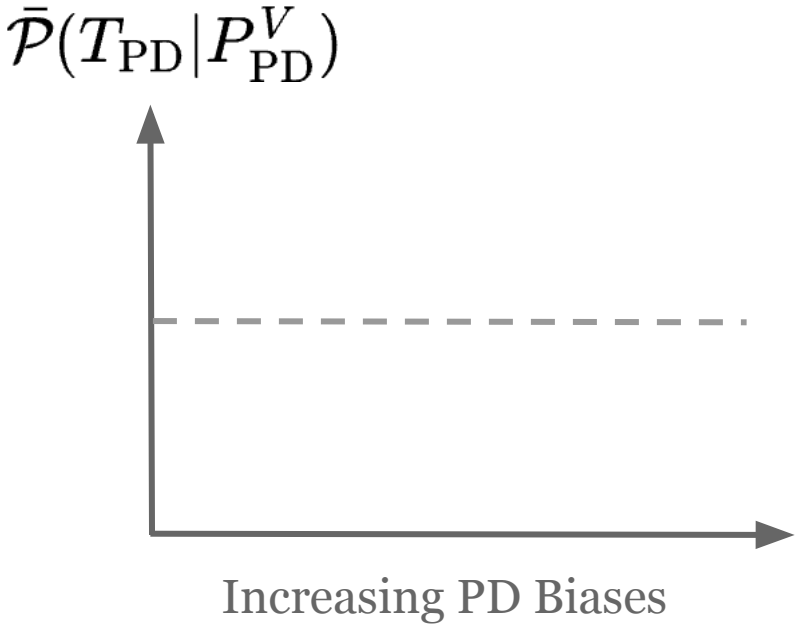
$$bias(V, \text{PD}) = \frac{1}{|\mathcal{S}_V|} \sum_{t_{\text{PD}} \in \mathcal{S}_V} \frac{\mathcal{P}(t_{\text{PD}})}{\mathcal{P}(t_{\text{PD}}) + \mathcal{P}(t_{\text{DO}})}$$

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

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

Predictions on the IFE

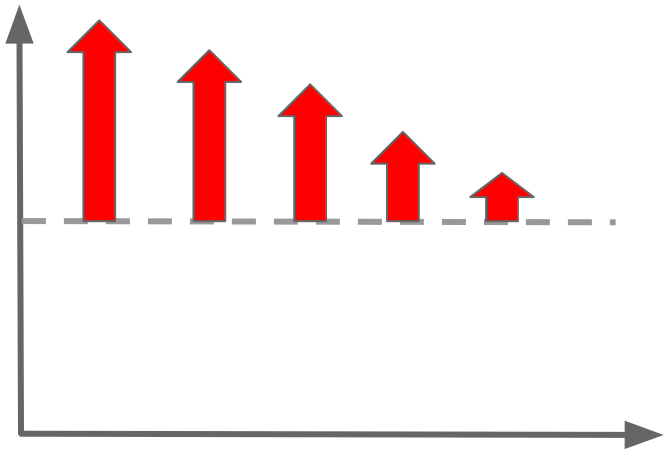
Simulations



Predictions on the IFE

Simulations

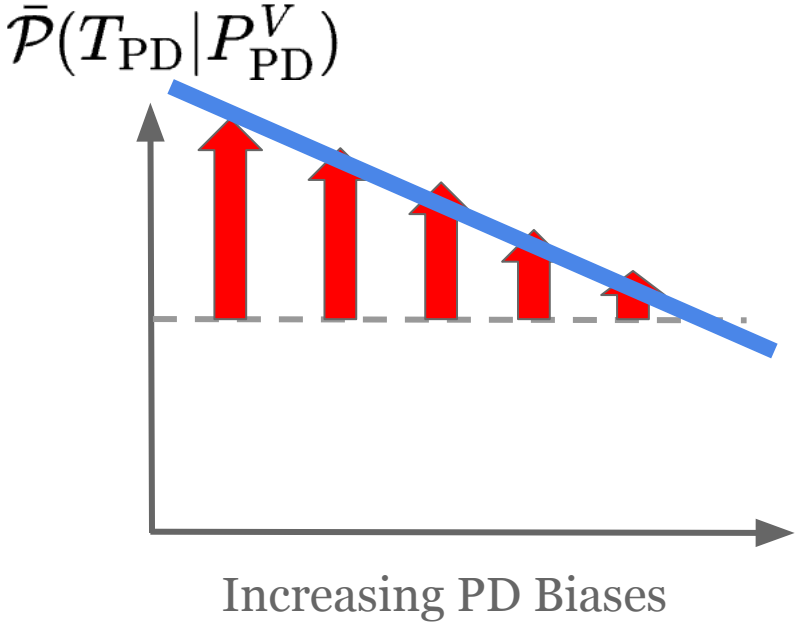
$\bar{\mathcal{P}}(T_{\mathrm{PD}}|P^{V}_{\mathrm{PD}})$



Increasing PD Biases

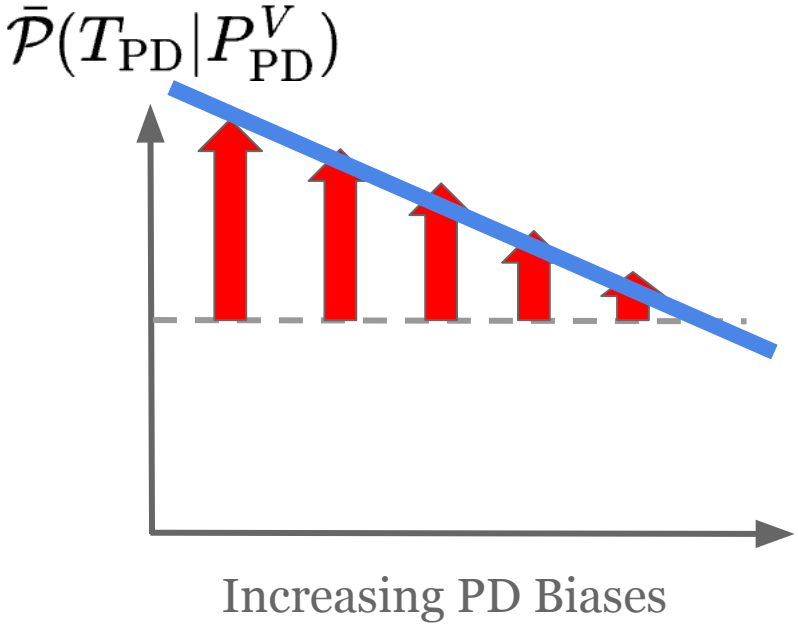
Predictions on the IFE

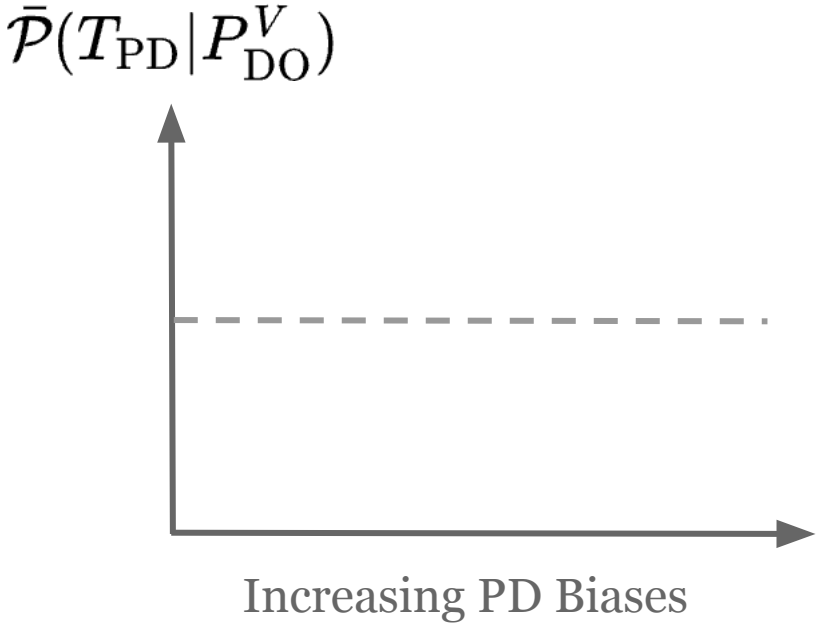
Simulations



Predictions on the IFE

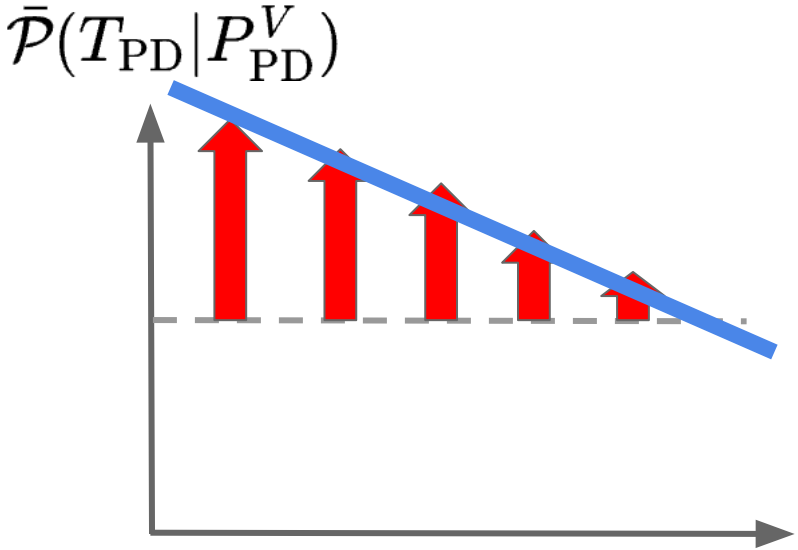
Simulations

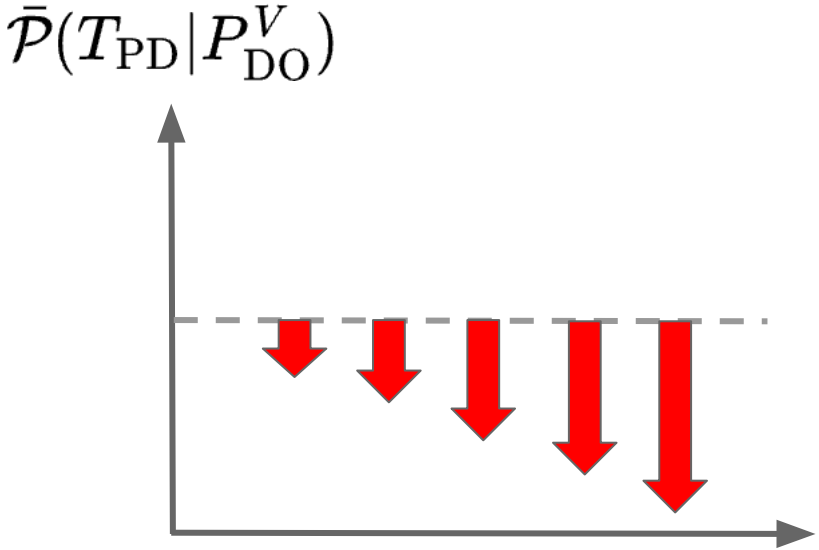




Predictions on the IFE

Simulations

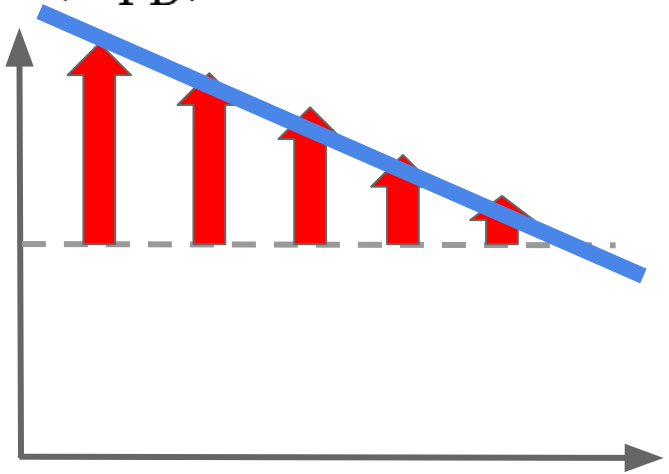




Predictions on the IFE

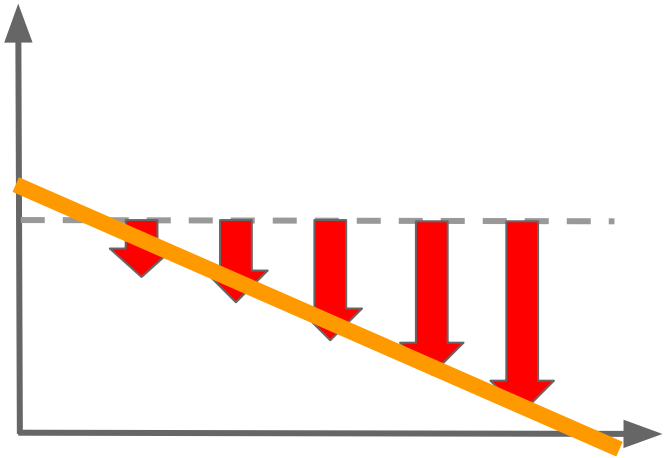
Simulations

$\bar{\mathcal{P}}(T_{\mathrm{PD}}|P^{V}_{\mathrm{PD}})$



Increasing PD Biases

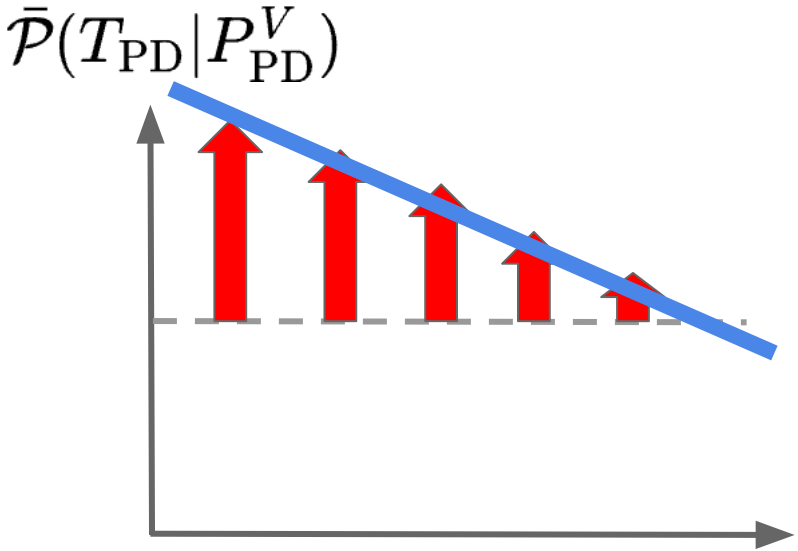
$\bar{\mathcal{P}}(T_{\mathrm{PD}}|P^{V}_{\mathrm{DO}})$



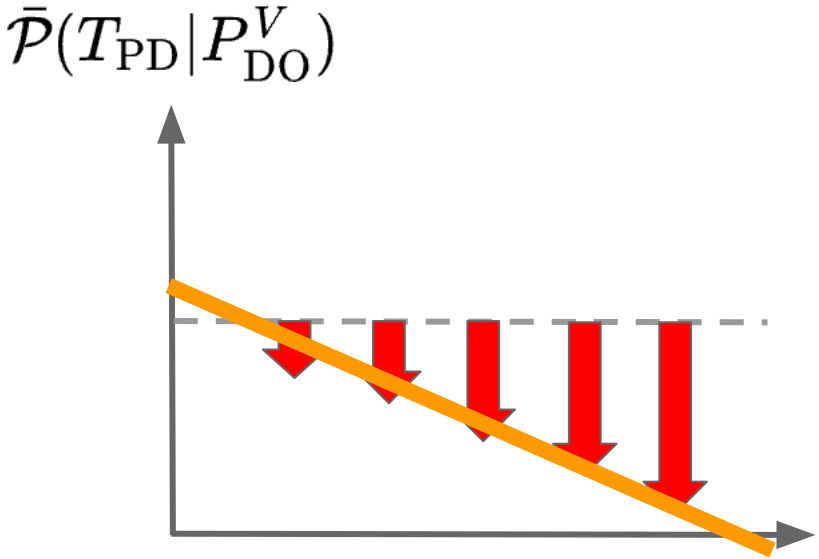
Increasing PD Biases

Predictions on the IFE

Simulations



Increasing PD Biases



Increasing PD Biases

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

Priming in LLMs

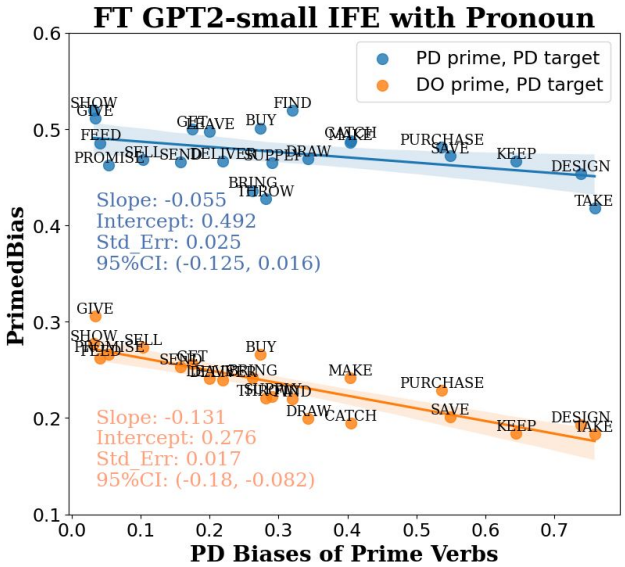
Results

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

Results

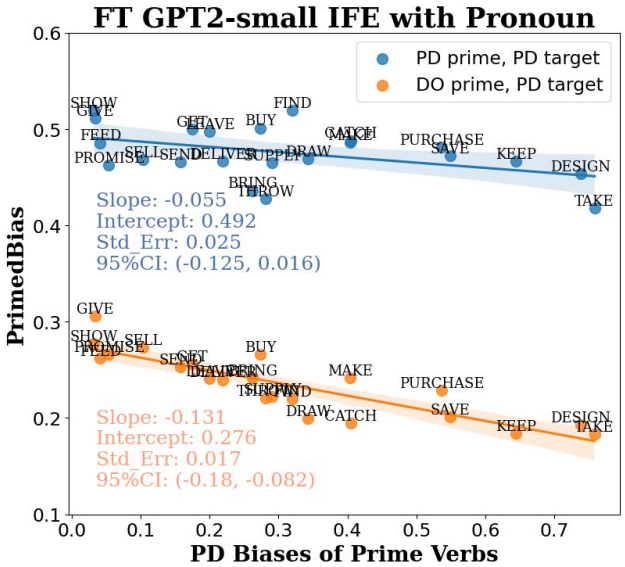
Fine-tuning Mode: fine-tuning the model with the prime sentence and use the updated model to run the target sentence — with weight update;



Priming in LLMs

Results

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

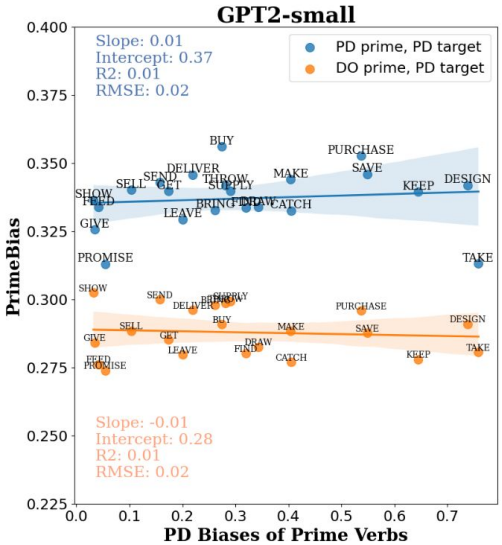
Results

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

Priming in LLMs

Results

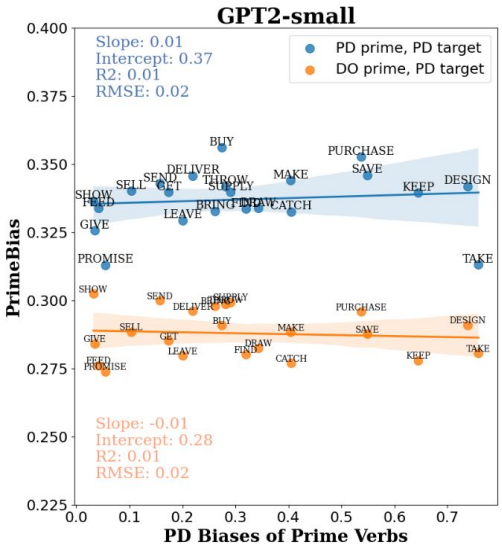
Concatenation Mode: concatenating the prime and target sentences and run the model — without weight update;



Priming in LLMs

Results

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

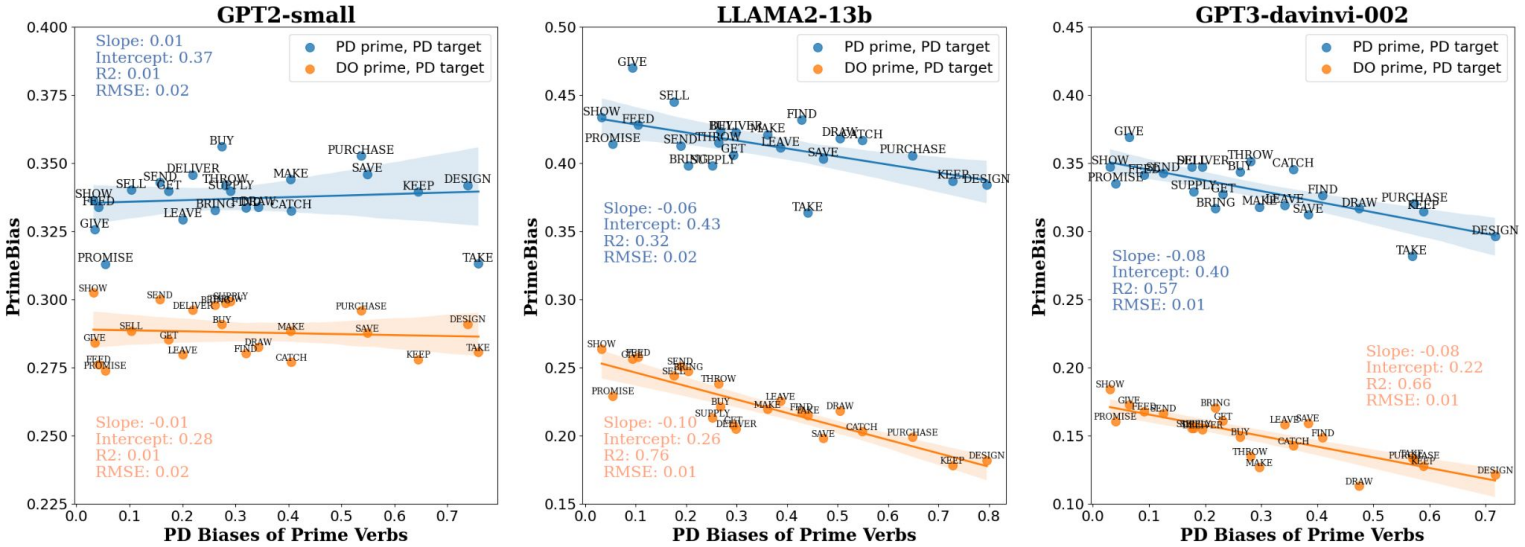


***Larger* models show more significant inverse frequency effects!**

Priming in LLMs

Results

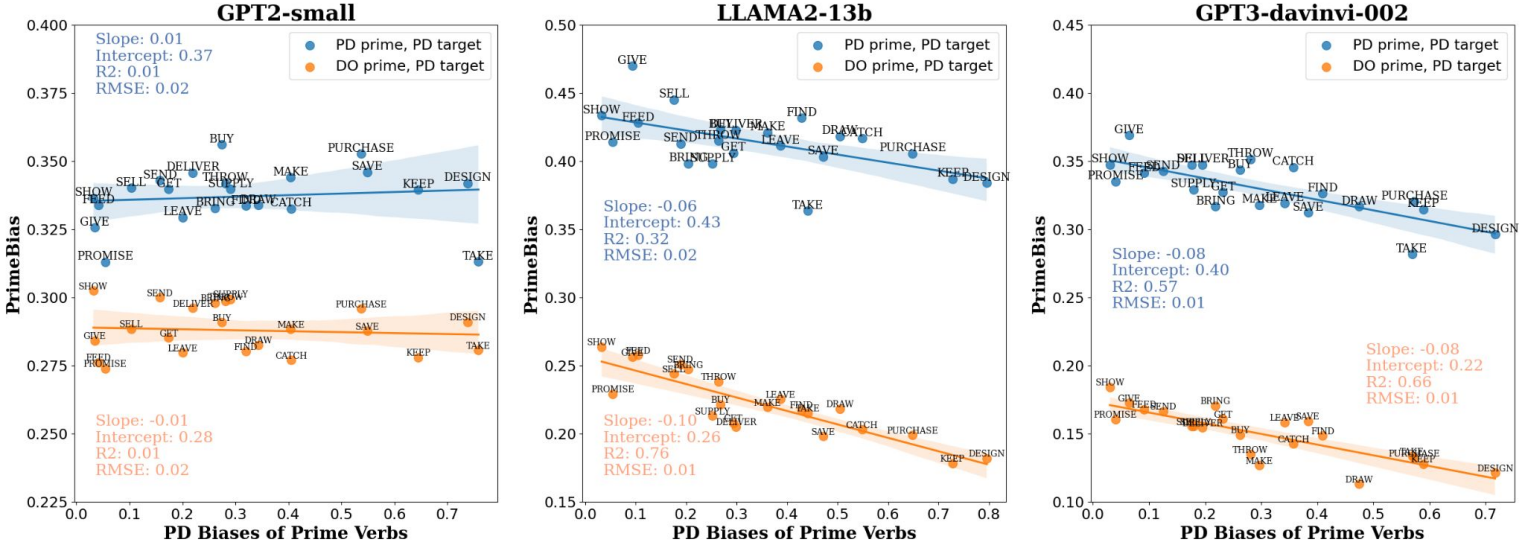
Concatenation Mode: concatenating the prime and target sentences and run the model — without weight update;



Priming in LLMs

Results

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



***Larger* models show more significant inverse frequency effects!**

Interim Summary 3

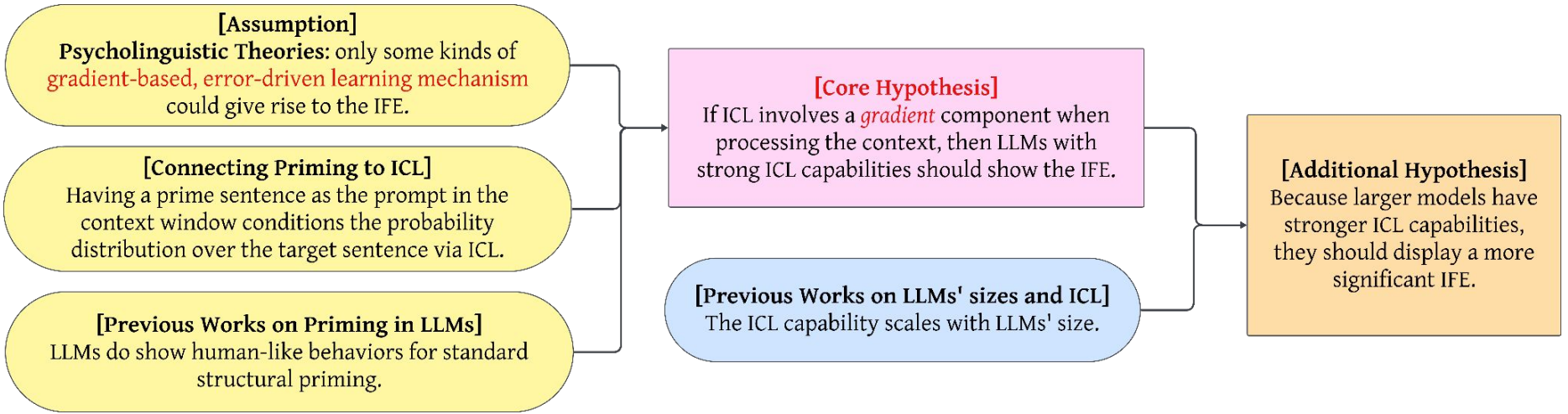
In-context Learning (ICL) in LLMs | Structural Priming | Current Study | Discussion & Implications

- *In-context* learning vs. *In-weights* learning
- ICL as a processing mechanism of LLMs;
- Is ICL functionally performing some error-driven learning?

- Structural priming effect;
- The Inverse Frequency Effect (IFE);
- Two accounts: transient activation vs. implicit learning;

- Analogy between priming and ICL;
- Do LLMs show the IFE? – yes, with various degrees!
- IFE as a diagnostic of error-driven learning!

Reasoning


[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

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

Takeaways

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

- Generalizing beyond standard notion of ICL, connecting priming with prompting 🌓

Takeaways

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

- Generalizing beyond standard notion of ICL, connecting priming with prompting 🌓
- Larger LLMs show *more significant* IFE 🤭🧐

Takeaways

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

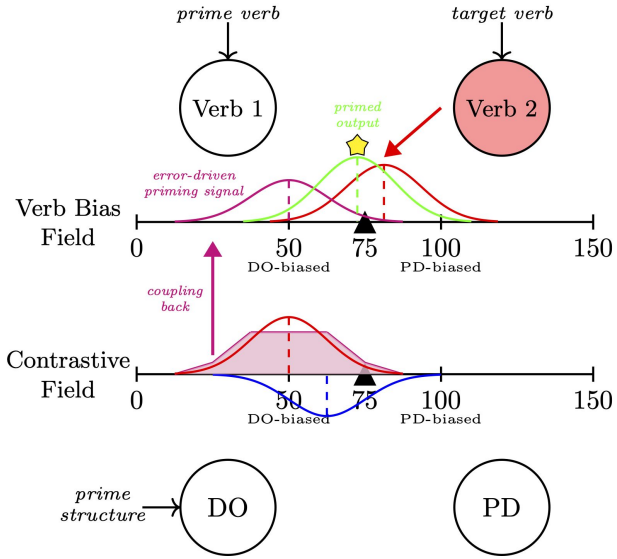
- Generalizing beyond standard notion of ICL, connecting priming with prompting 🌓
- Larger LLMs show *more significant* IFE 🤭🧐
- At least in the case of priming, error-driven learning happens in ICL, supporting the hypothesis that [LLMs is functionally performing gradient descent] 🌟🌟🌟

Thanks for Listening!
Q&A Session 🙋‍♂️🙋‍♀️

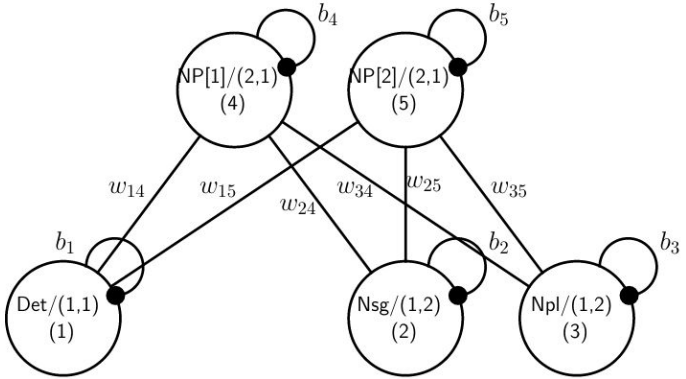
In-context Learning (ICL) in LLMs	Structural Priming	Current Study	Discussion & Implications
• *In-context* learning vs. *In-weights* learning • ICL as a processing mechanism of LLMs; • ICL is functionally equivalent to Gradient Descent?	• Structural priming effect; • The Inverse Frequency Effect (IFE); • Two accounts: transient activation vs. implicit learning;	• Analogy between priming and ICL? • Do LLMs show the IFE? • IFE as a diagnostic of error-driven learning?	• Larger models show stronger IFE; • There is an *implicit* gradient component involved in ICL; • Humans and LLMs share a similar processing mechanism!

Open Discussion: activation vs. learning?

Enriching Transient Activation to simulate implicit learning?

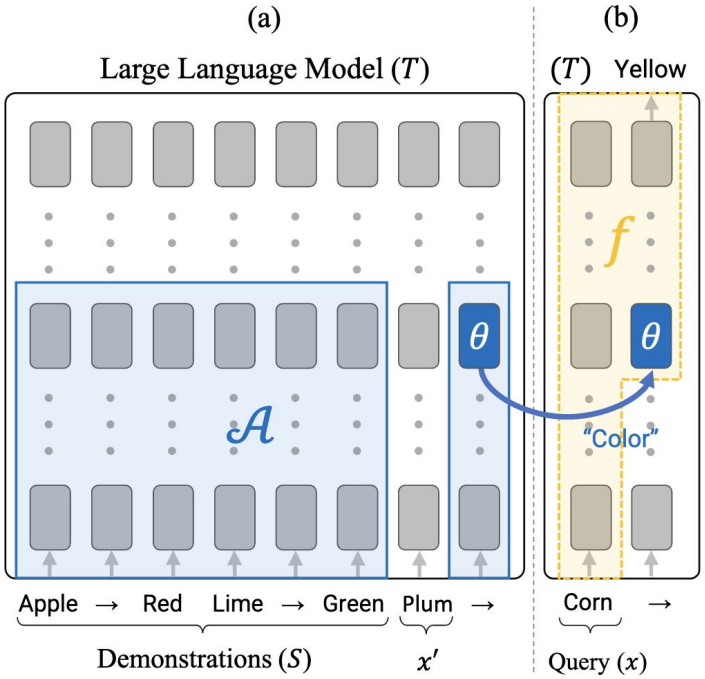


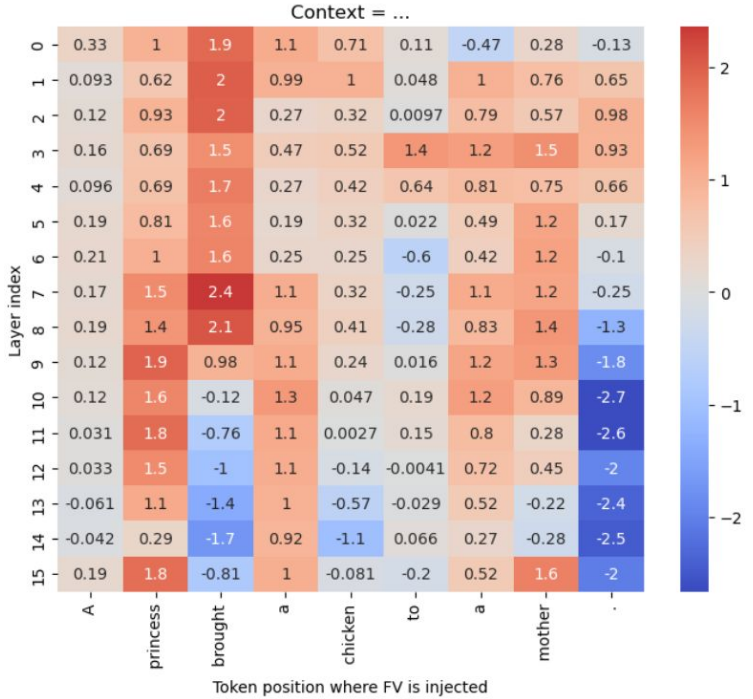
Dynamic Field Theory



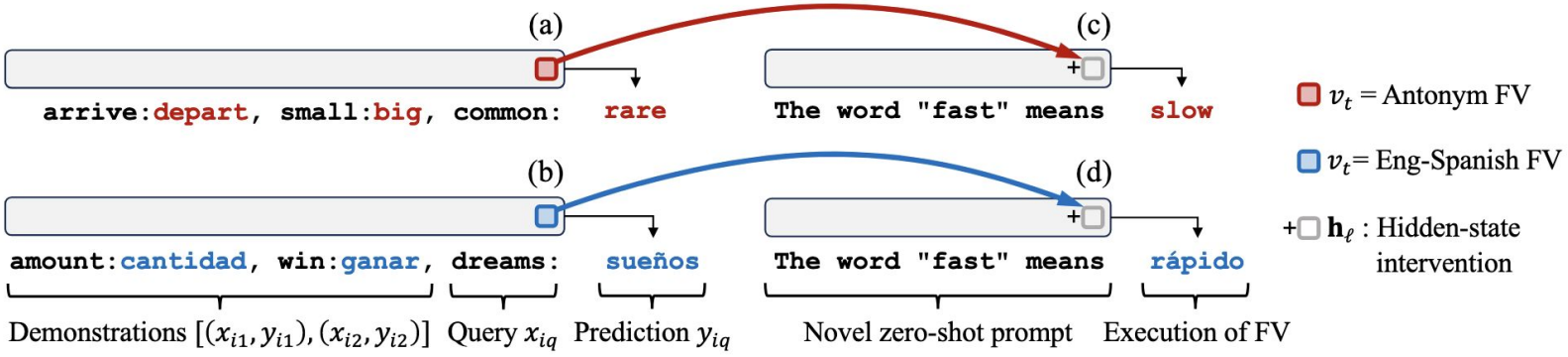
Gradient Symbolic Computation

On-going Work 1: Task / Function Vectors?



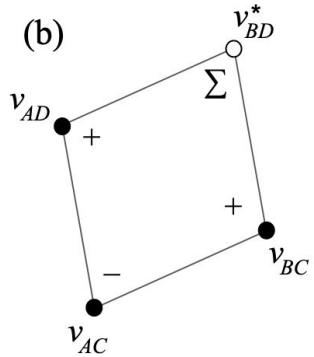


On-going Work 1: Task / Function Vectors cont.



(a) Input: `"Italy, Russia, China, Japan, France"`

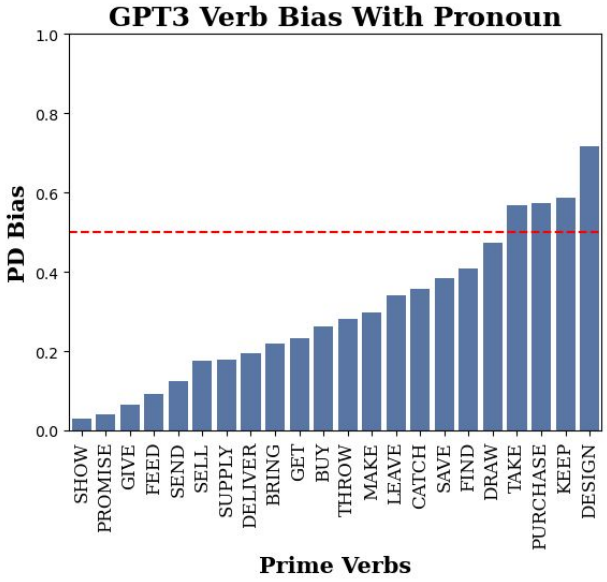
FV	Task	Expected Output
v_{AC}	First-Copy	`Italy`
v_{AD}	First-Capital	`Rome`
v_{BC}	Last-Copy	`France`
v^*_{BD}	Last-Capital	`Paris`

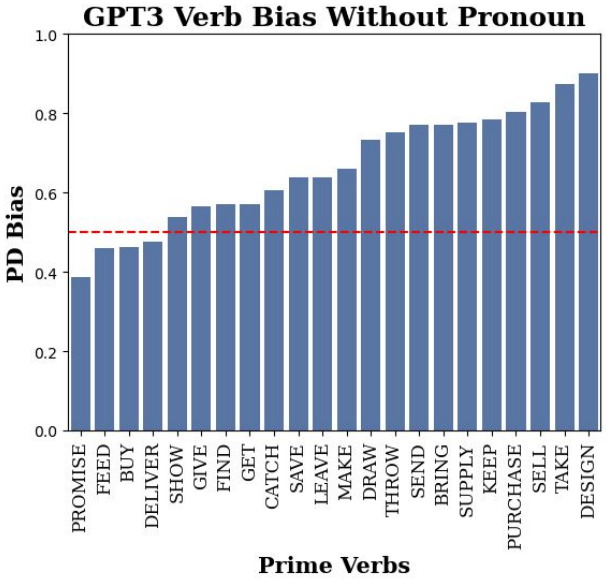


On-going Work 2: IFE in other ICL “Tasks”?

- If the IFE is a diagnostic for the error-driven nature of adaptation, then we could apply it to any other ICL tasks;
  1. **Other structural priming instances:**
     a. Active-passive;
     b. Complementizer (*that*) - priming;
  2. **Non-linguistic tasks where we can sort demonstration difficulty w.r.t baseline performance;**
     a. Two-digit {addition, multiplication}: 10 x 10 is easier than 31 x 67;
     b. Country-capital mapping: more commonly known <country, capital> pairs leads to lower surprisal than less known pairs;

More Intricate Pattern: Pronoun vs. NoPronoun





Pronoun Corpus: replacing all indirect object with a pronoun;

Pronoun vs. NoPronoun: results

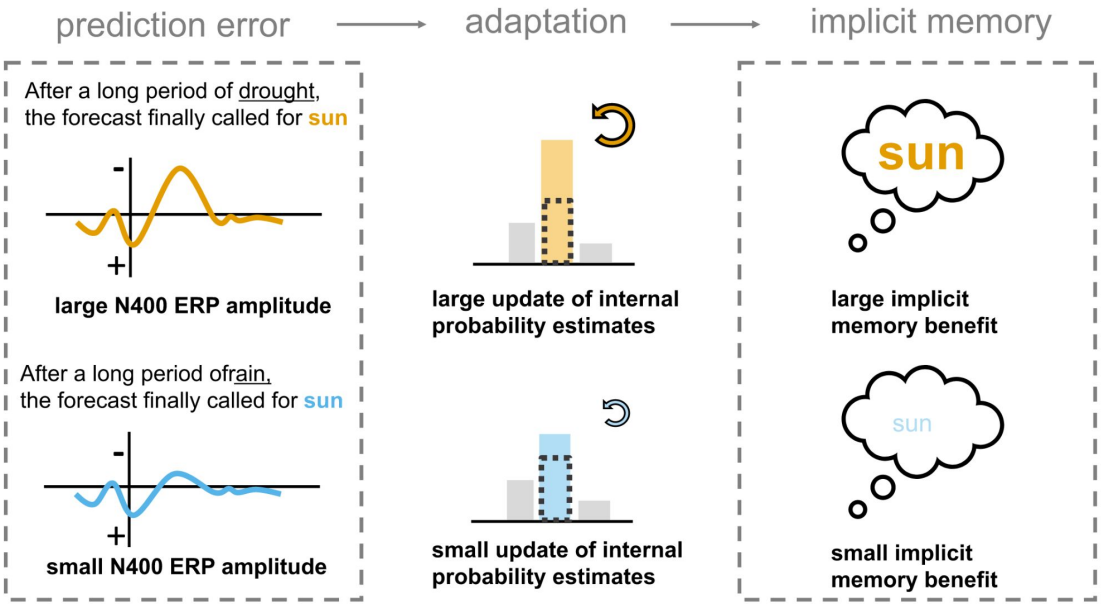
		Intercept		Slope		95% CI of Slopes	
Model	Pronoun Obj?	PD-PD	DO-PD	PD-PD	DO-PD	PD-PD	DO-PD
GPT2-small	True	0.370	0.278	0.011	-0.007	(-0.049, 0.072)	(-0.058, 0.044)
GPT2-small	False	0.746	0.653	0.014	0.006	(-0.042, 0.07)	(-0.06, 0.072)
GPT2-medium	True	0.351	0.256	-0.013	-0.026	(-0.078, 0.053)	(-0.073, 0.022)
GPT2-medium	False	0.748	0.590	-0.023	-0.035	(-0.079, 0.032)	(-0.125, 0.054)
GPT2-large	True	0.330	0.241	0.011	-0.037^{+}	(-0.043, 0.065)	(-0.09, 0.015)
GPT2-large	False	0.698	0.487	-0.003	-0.02	(-0.062, 0.055)	(-0.098, 0.058)
Llama-7b	True	0.392	0.229	-0.02	-0.086****	(-0.067, 0.026)	**(-0.126, -0.045)**
Llama-7b	False	0.807	0.627	-0.026	-0.111^{+}	(-0.102, 0.049)	(-0.279, 0.057)
Llama-7b-chat	True	0.413	0.263	-0.012	-0.095****	(-0.067, 0.043)	**(-0.146, -0.044)**
Llama-7b-chat	False	0.788	0.605	-0.013	-0.102	(-0.115, 0.089)	(-0.289, 0.085)
Llama-13b	True	0.434	0.256	-0.059**	-0.099****	**(-0.113, -0.005)**	**(-0.134, -0.063)**
Llama-13b	False	0.859	0.685	-0.066^{+}	-0.177*	(-0.163, 0.031)	(-0.388, 0.033)
GPT3-davinci-002	True	0.403	0.223	-0.078****	-0.078****	**(-0.121, -0.035)**	**(-0.114, -0.043)**
GPT3-davinci-002	False	0.851	0.632	-0.064^{+}	-0.145*	(-0.153, 0.025)	(-0.301, 0.012)

Pattern: more significant IFE results in the WithPronoun condition than the NoPronoun condition;

Interpretation: ICL capability correlates with data scale – more WithPronoun sentences in the training data!

Future Direction: ERP as evidence of learning signal?

The N400 ERP component reflects an error-based implicit learning signal during language comprehension



Predictive coding:
A theory under which the functional role of neural systems centers on prediction and learning from prediction error.

Future Direction: ERP as evidence of learning signal, cont.

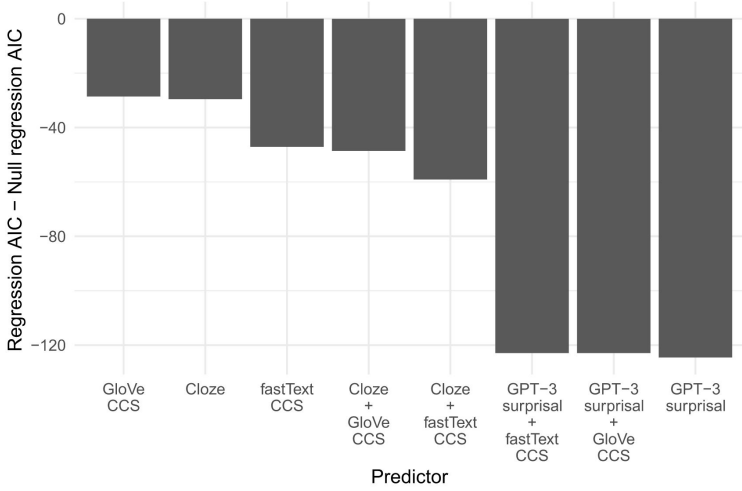
SPECIAL ISSUE:
Cognitive Computational Neuroscience of Language

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



Neural language model gradients predict event-related brain potentials

Stefan L. Frank
Centre for Language Studies
Radboud University
Nijmegen, the Netherlands
stefan.frank@ru.nl

Additional Slides

LLMs as Psycholinguistic Subjects

Model

Existing Line of research: treating LMs as psycholinguistic subjects and probing whether they have acquired the linguistic knowledge that humans used for sentence processing.

LLMs as Psycholinguistic Subjects

Model

Existing Line of research: treating LMs as psycholinguistic subjects and probing whether they have acquired the linguistic knowledge that humans used for sentence processing.

The Emergence of Number and Syntax Units in LSTM Language Models

Yair Lakretz
Cognitive Neuroimaging Unit
NeuroSpin center
91191, Gif-sur-Yvette, France
yair.lakretz@gmail.com

German Kruszewski
Facebook AI Research
Paris, France
germank@gmail.com

Theo Desbordes
Facebook AI Research
Paris, France
tdesbordes@fb.com

Dieuwke Hupkes
ILLC, University of Amsterdam
Amsterdam, Netherlands
d.hupkes@uva.nl

Stanislas Dehaene
Cognitive Neuroimaging Unit
NeuroSpin center
91191, Gif-sur-Yvette, France
stanislas.dehaene@gmail.com

Marco Baroni
Facebook AI Research
Paris, France
mbaroni@fb.com

LLMs as Psycholinguistic Subjects

Model

Existing Line of research: treating LMs as psycholinguistic subjects and probing whether they have acquired the linguistic knowledge that humans used for sentence processing.

What do RNN Language Models Learn about Filler–Gap Dependencies?

Ethan Wilcox[1], **Roger Levy**[2], **Takashi Morita**[3,4], and **Richard Futrell**[5]

[1]Department of Linguistics, Harvard University, wilcoxeg@g.harvard.edu
[2]Department of Brain and Cognitive Sciences, MIT, rplevy@mit.edu
[3]Primate Research Institute, Kyoto University, tmorita@alum.mit.edu
[4]Department of Linguistics and Philosophy, MIT
[5]Department of Language Science, UC Irvine, rfutrell@uci.edu

LLMs as Psycholinguistic Subjects

Model

Existing Line of research: treating LMs as psycholinguistic subjects and probing whether they have acquired the linguistic knowledge that humans used for sentence processing.

When a sentence does not introduce a discourse entity, Transformer-based models still sometimes refer to it

Sebastian Schuster
Center for Data Science
Department of Linguistics
New York University
schuster@nyu.edu

Tal Linzen
Center for Data Science
Department of Linguistics
New York University
linzen@nyu.edu

LLMs as Psycholinguistic Subjects

Model

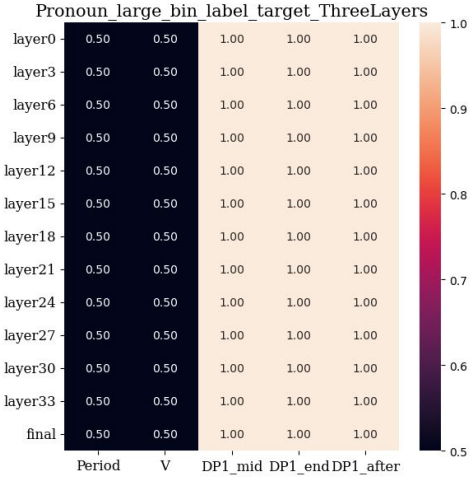
Existing Line of research: treating LMs as psycholinguistic subjects and probing whether they have acquired the linguistic knowledge that humans used for sentence processing.

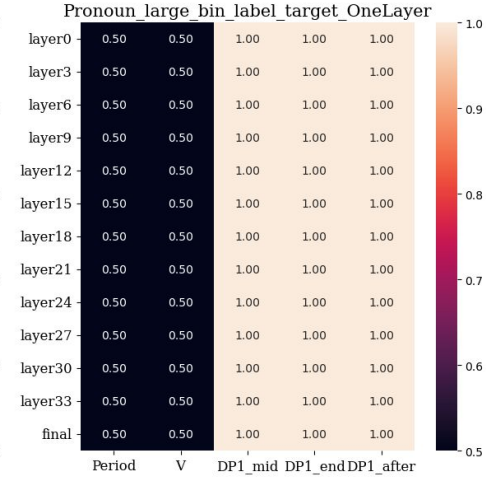
Neural Language Models as Psycholinguistic Subjects: Representations of Syntactic State

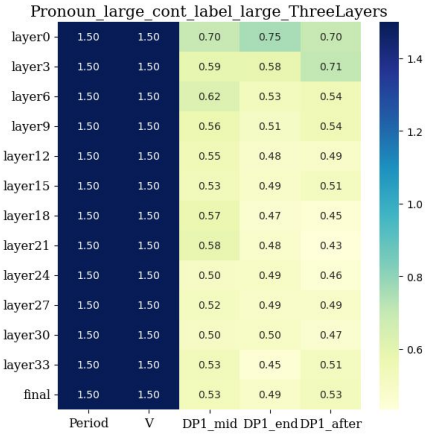
Richard Futrell[1], **Ethan Wilcox**[2], **Takashi Morita**[3,4], **Peng Qian**[5], **Miguel Ballesteros**[6], and **Roger Levy**[5]

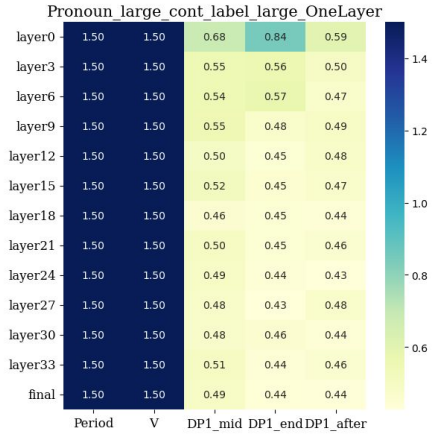
[1]Department of Language Science, UC Irvine, rfutrell@uci.edu
[2]Department of Linguistics, Harvard University, wilcoxeg@g.harvard.edu
[3]Primate Research Institute, Kyoto University, tmorita@alum.mit.edu
[4]Department of Linguistics and Philosophy, MIT
[5]Department of Brain and Cognitive Sciences, MIT, {pqian,rplevy}@mit.edu
[6]IBM Research, MIT-IBM Watson AI Lab, miguel.ballesteros@ibm.com

Probing Classifiers

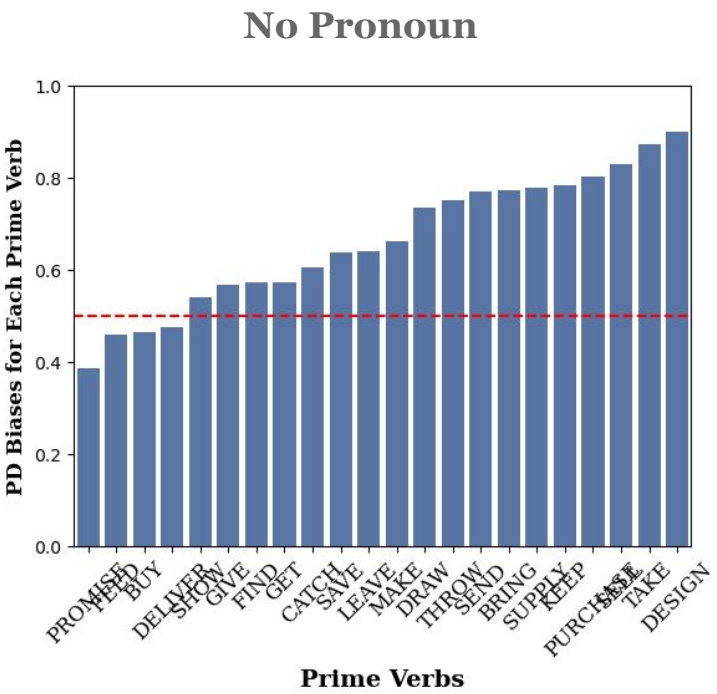


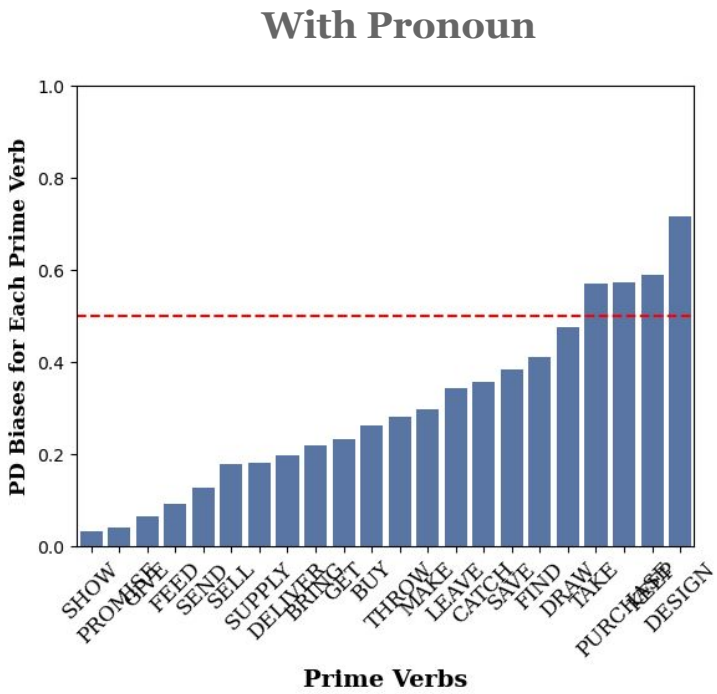




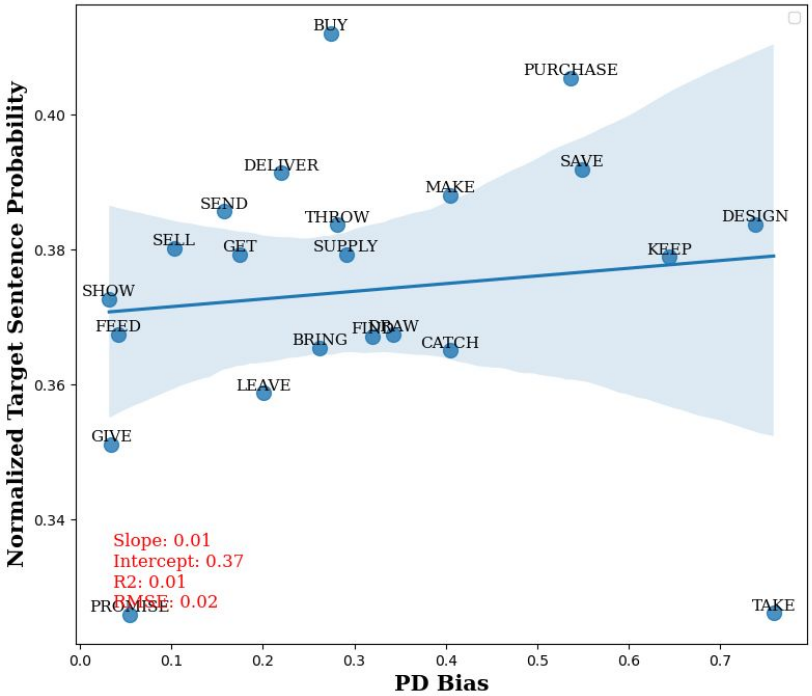


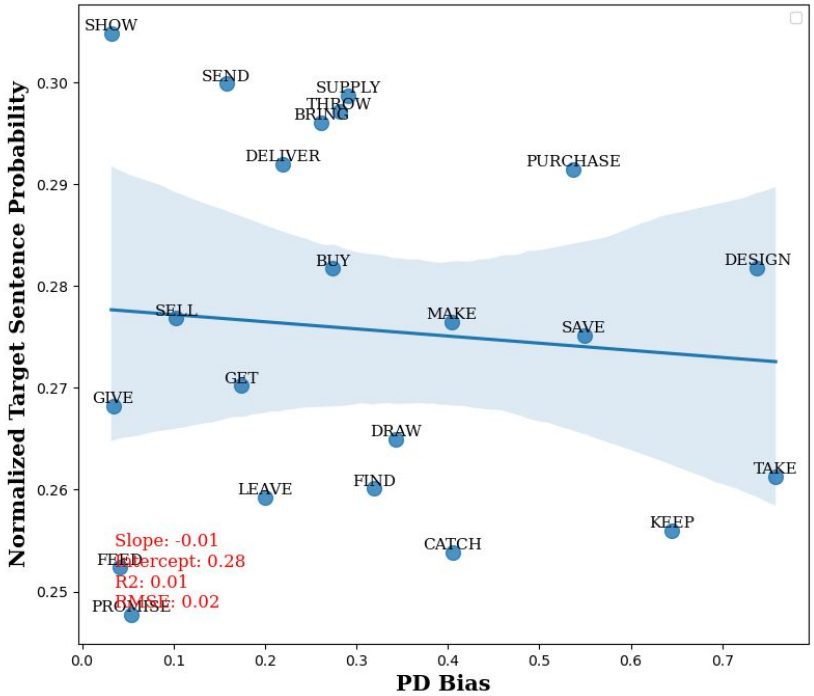
Verb Bias: With Pronoun vs. No Pronoun



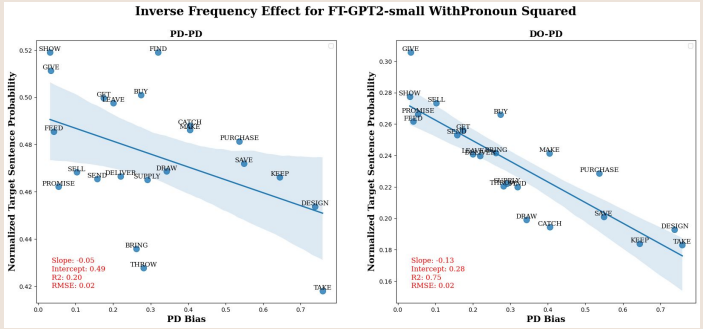


GPT2-small doesn't show IFE in TA mode





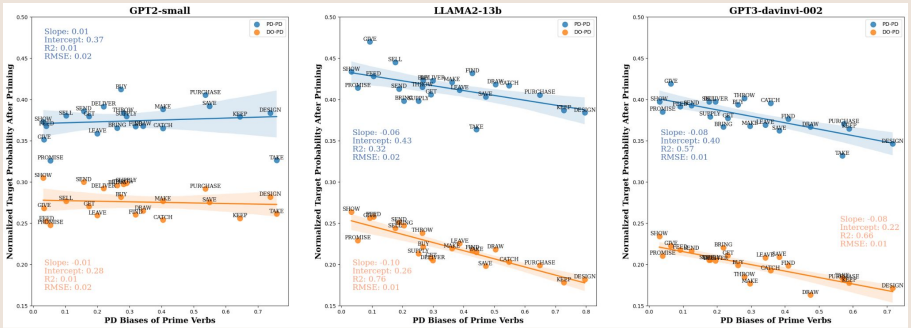
Explanations?



Implicit Learning Mode: GPT2-small does have the capability to capture the IFE!

Presence vs. absence of Error-driven weight update mechanism!

There is some mechanism that functionally performs gradient update in LLMs!



Transient Activation Mode: GPT2-small doesn't show the IFE, while larger models show stronger IFEs.

In-Context Learning

Thanks for Listening!
Q&A Session 🤦

In-context Learning (ICL) in LLMs	Structural Priming	Current Study	Discussion & Implications
• *In-context* learning vs. *In-weights* learning • ICL as a processing mechanism of LLMs; • Is ICL functionally performing some error-driven learning?	• Structural priming effect; • The Inverse Frequency Effect (IFE); • Two accounts: transient activation vs. implicit learning;	• Analogy between priming and ICL? • Do LLMs show the IFE? • IFE as a diagnostic of error-driven learning?	• Larger models show stronger IFE; • There is an *implicit* gradient component involved in ICL; • Humans and LLMs share a similar processing mechanism!

Selected Reference

- Sarah Bernolet and Robert J. Hartsuiker. 2010. Does verb bias modulate syntactic priming? *Cognition*, 114(3):455–461.
- J. Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive Psychology*, 18(3):355– 387.
- Kathryn Bock and Carol A Miller. 1991. Broken agreement. *Cognitive Psychology*, 23(1):45–93.
- Laurel Brehm, Pyeong Whan Cho, Paul Smolensky, and Matthew A. Goldrick. 2022. PIPS: A Parallel Planning Model of Sentence Production. *Cognitive Science*, 46(2):e13079.
- Pyeong Whan Cho, Matthew Goldrick, Richard L. Lewis, and Paul Smolensky. 2018. Dynamic encoding of structural uncertainty in gradient symbols. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 19–28, Salt Lake City, Utah. Association for Computational Linguistics.
- Pyeong Whan Cho, Matthew Goldrick, and Paul Smolensky. 2020. Parallel parsing in a Gradient Symbolic Computation parser.
- John Hale and Paul Smolensky. 2006. Harmonic gram- mars and harmonic parsers for formal languages. *Smolensky and Legendre*, pages 393–416.
- T. Florian Jaeger and Neal Snider. 2007. Implicit Learning and Syntactic Persistence: Surprisal and Cumulativity. *University of Rochester Working Papers in the Language Sciences*, 3:26–44.
- Martin J. Pickering and Holly P. Branigan. 1998. The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, 39(4):633–651.
- Eunkyung Yi, Jean-Pierre Koenig, and Douglas Roland. 2019. Semantic similarity to high-frequency verbs affects syntactic frame selection. *Cognitive Linguistics*, 30(3):601–628.