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

Zhenghao “Herbert” Zhou, Robert Frank, R. Thomas McCoy

Yale Linguistics
NAACL 2025

Evidence from the Inverse Frequency Effects in Structural Priming

Outline

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);
- Implicit learning, a type of error-driven learning, accounts for the IFE;;

Outline

In-context Learning (ICL) in LLMs	Structural Priming	Current Study
• *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); • Implicit learning, a type of error-driven learning, accounts for the IFE;;	• 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);
- Implicit learning, a type of error-driven learning, accounts for the IFE;;

Current Study

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

Discussion & Implications

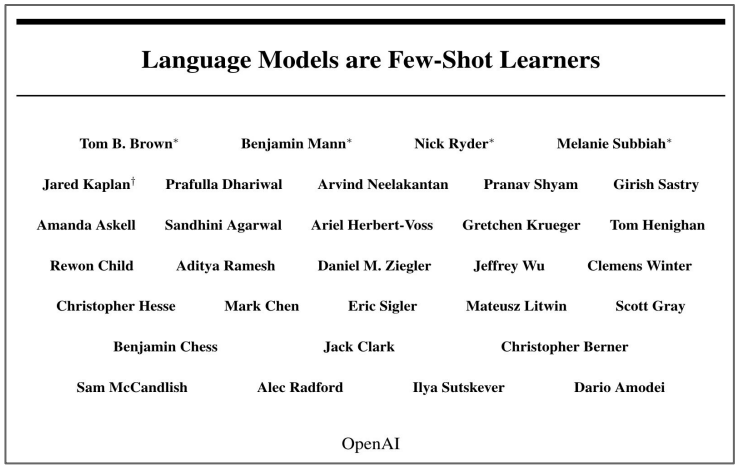
- Larger models show stronger IFE;
- There is an *implicit* error term involved in ICL;
- Humans and LLMs share a similar processing mechanism;

In-context Learning in LLMs

[Brown et al. 2020]

In-context Learning in LLMs

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

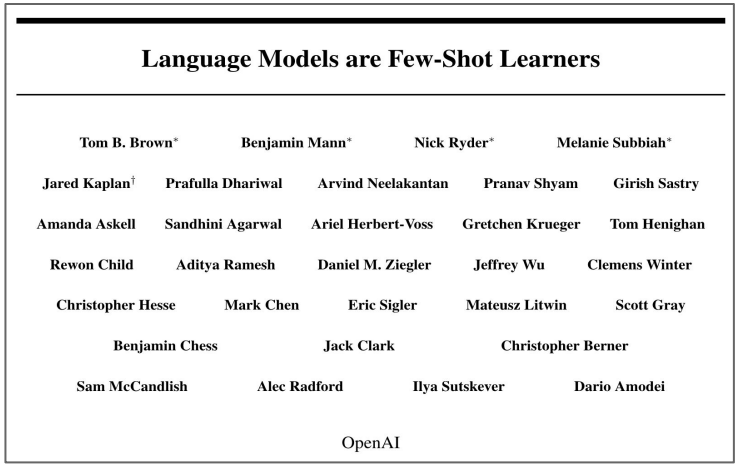
Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

[Brown et al. 2020]

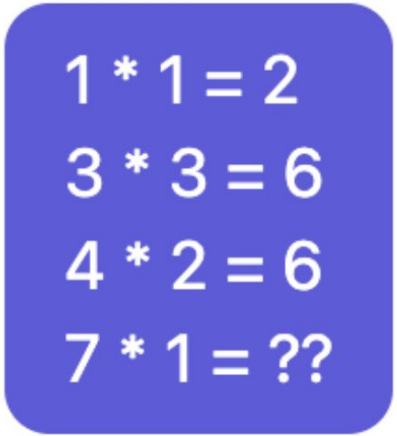
In-context Learning in LLMs

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

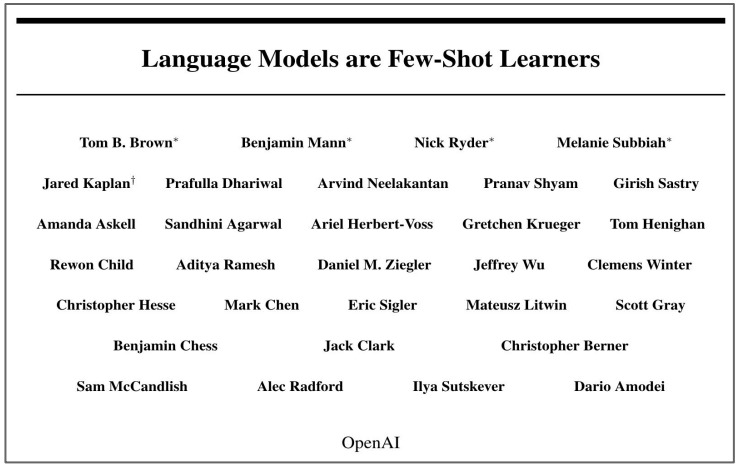
Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*
Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry
Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan
Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter
Christopher Hesse 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]

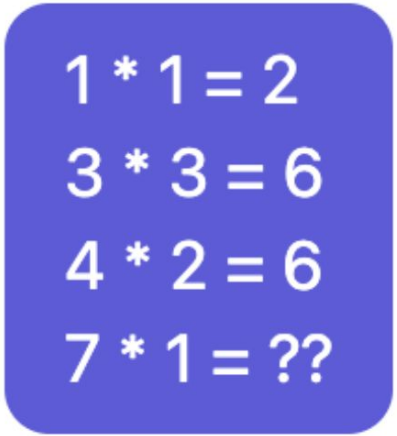
In-context Learning in LLMs

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



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

[Brown et al. 2020]

Open Question: ICL ≈(functional) Gradient Descent?

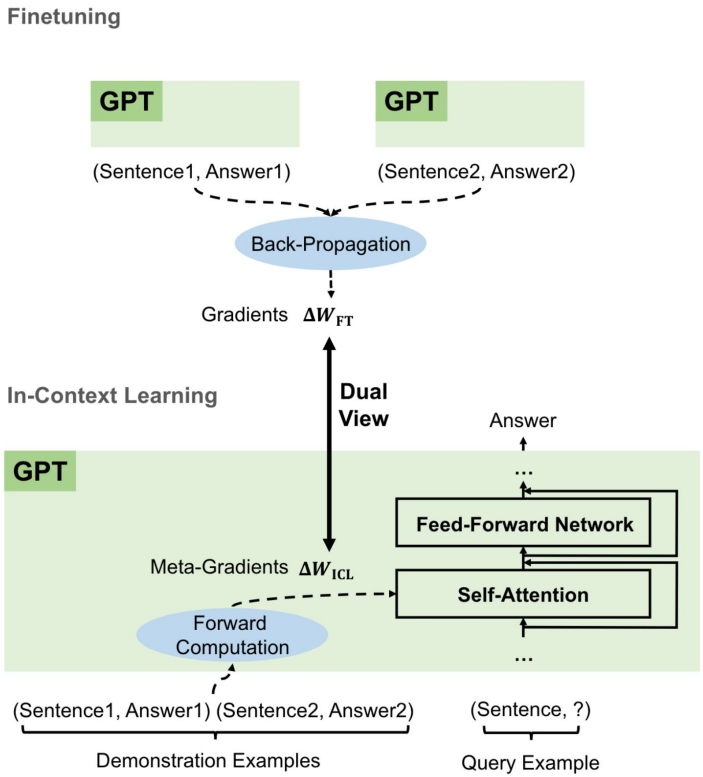
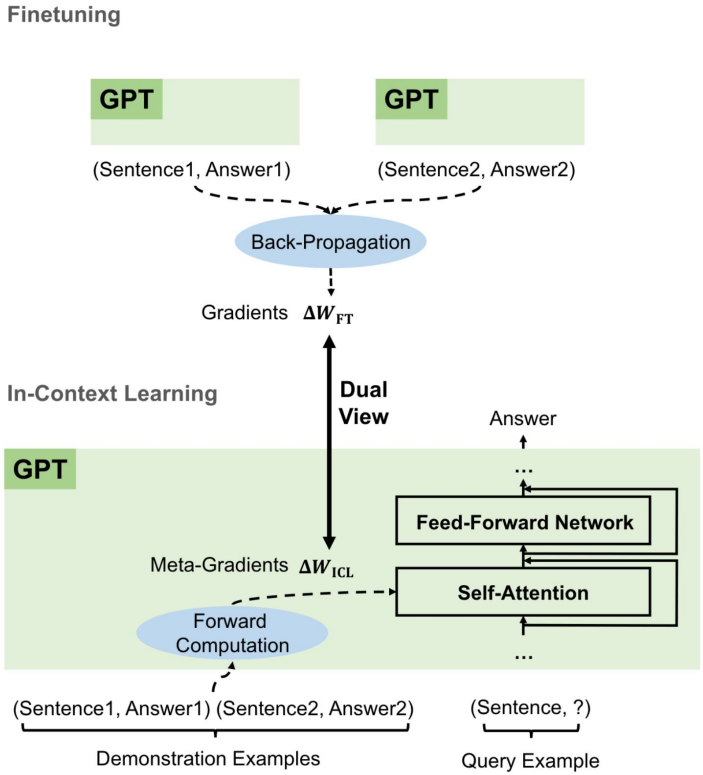


Figure reprinted from Dai et al. 2023;
E.g. [Xie et al. 2022, von Oswald et al. 2022]

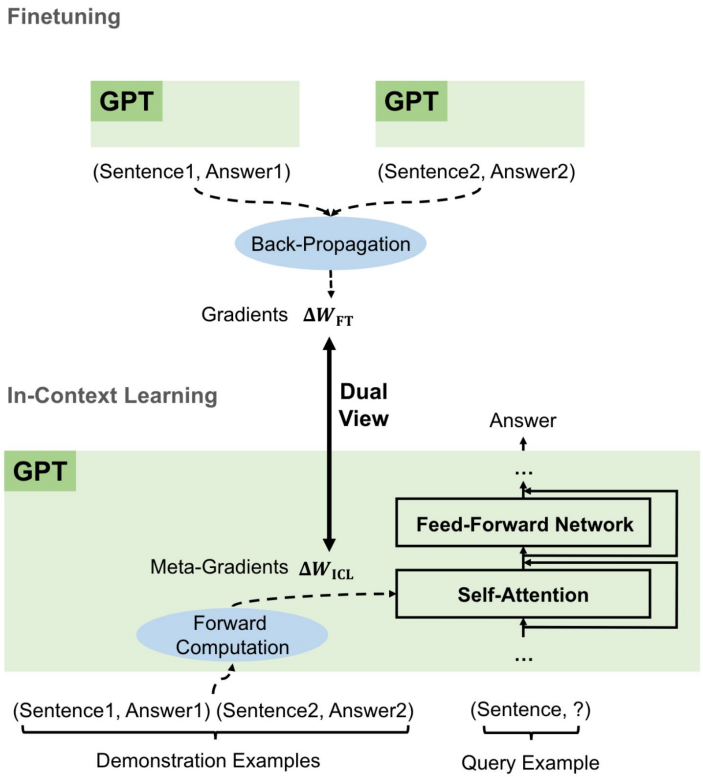
Open Question: ICL ≈(functional) Gradient Descent?



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

Figure reprinted from Dai et al. 2023;
E.g. [Xie et al. 2022, von Oswald et al. 2022]

Open Question: ICL ≈(functional) Gradient Descent?

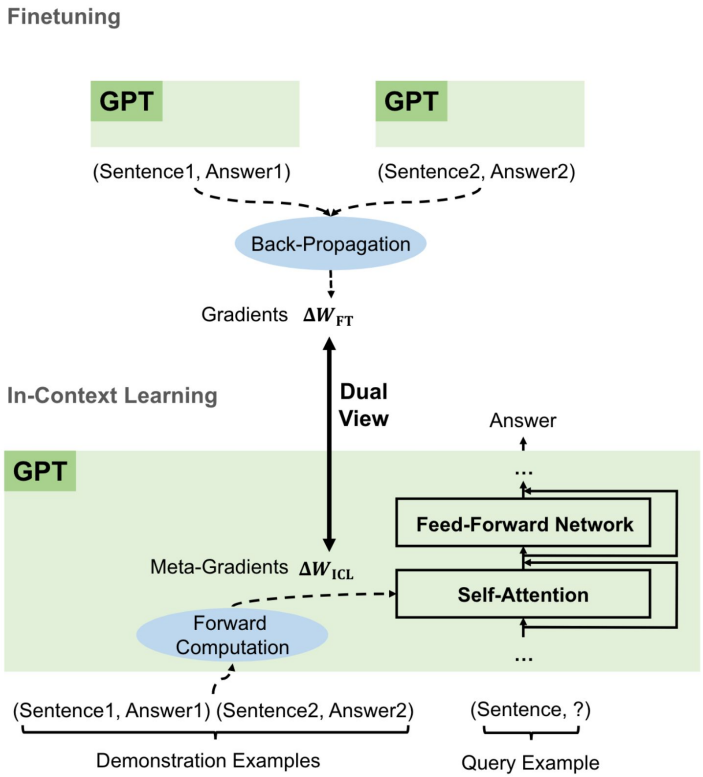


In-context Learning as implicitly performing gradient descent? — *in principle*, yes…

- **ICL performs implicit Bayesian inference;**

Figure reprinted from Dai et al. 2023;
E.g. [Xie et al. 2022, von Oswald et al. 2022]

Open Question: ICL ≈(functional) Gradient Descent?

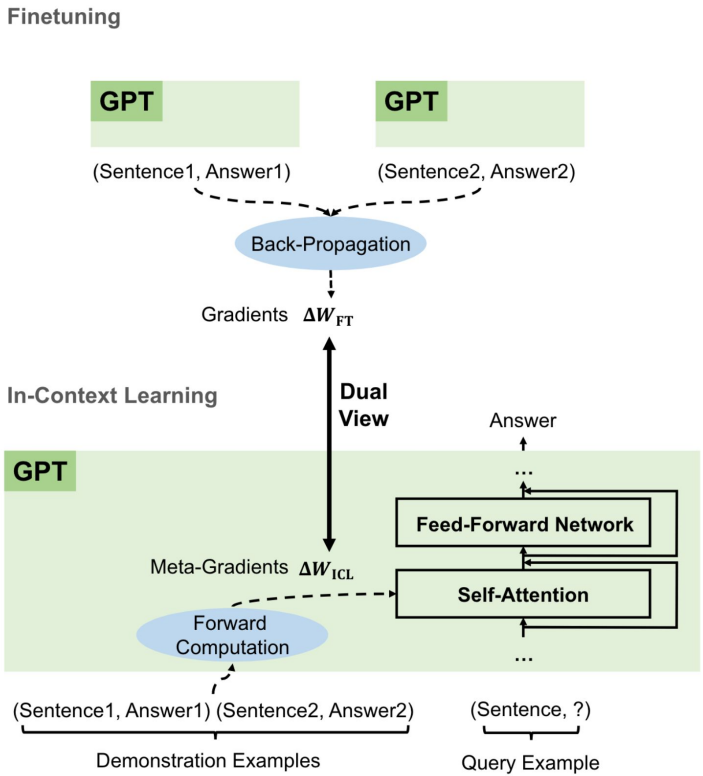


In-context Learning as implicitly performing gradient descent? — *in principle*, yes…

- **ICL performs implicit Bayesian inference;**
- **ICL functionally performs gradient descent;**

Figure reprinted from Dai et al. 2023;
E.g. [Xie et al. 2022, von Oswald et al. 2022]

Open Question: ICL ≈(functional) Gradient Descent?

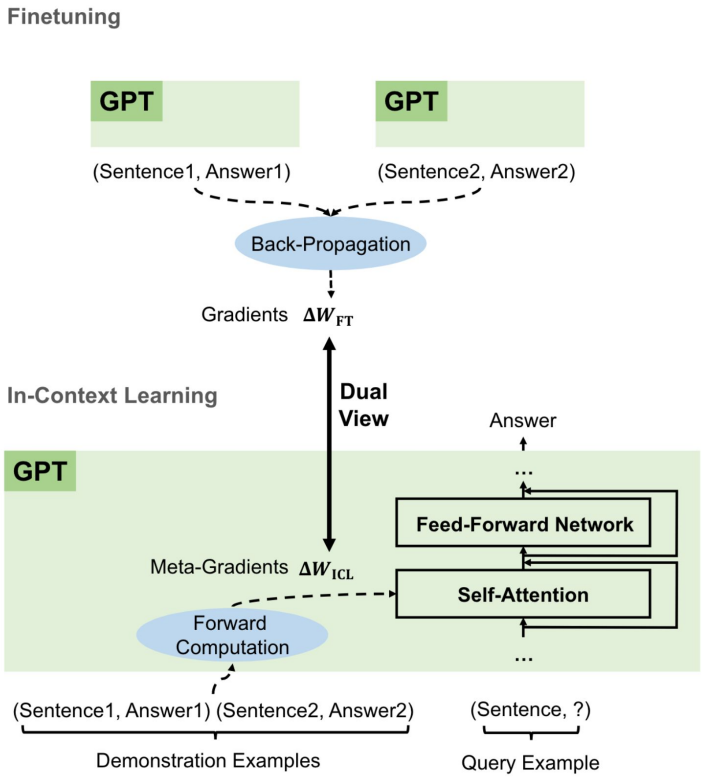


In-context Learning as implicitly performing gradient descent? — *in principle*, yes…

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

Figure reprinted from Dai et al. 2023;
E.g. [Xie et al. 2022, von Oswald et al. 2022]

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!

Figure reprinted from Dai et al. 2023;
E.g. [Xie et al. 2022, von Oswald et al. 2022]

Structural Priming

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]

Structural Priming

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

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*.

E.g. [Bock 1986, Chang 2012]

Structural Priming

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]

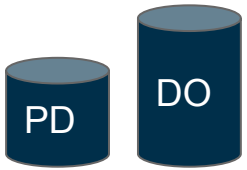
Inverse Frequency Effect (IFE)

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)

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

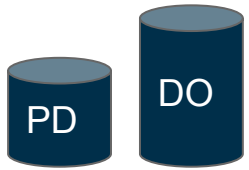


Verb Bias:
buy is biased towards DO
design towards PD

E.g. [Jaeger & Snider 2007]

Inverse Frequency Effect (IFE)

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

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

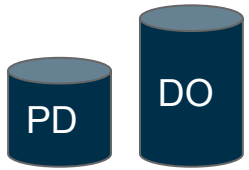
Prime in DO Structure **Target in PD Structure**

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)

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

$T_{PD} | P_{DO}$

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]

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*

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!

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

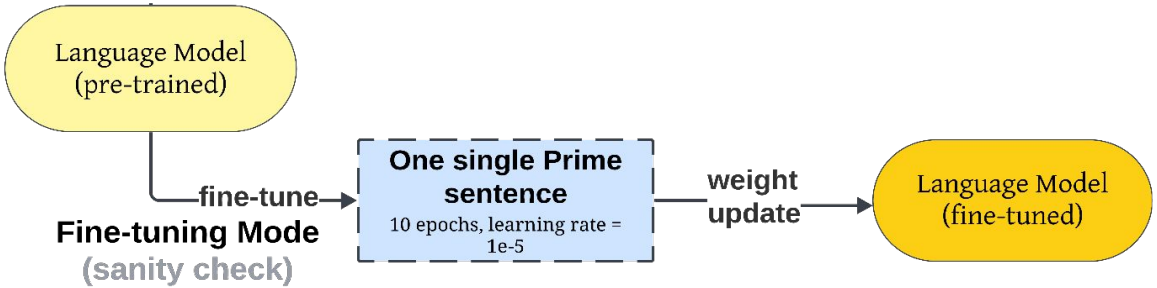
A terrible movie. → *negative*

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

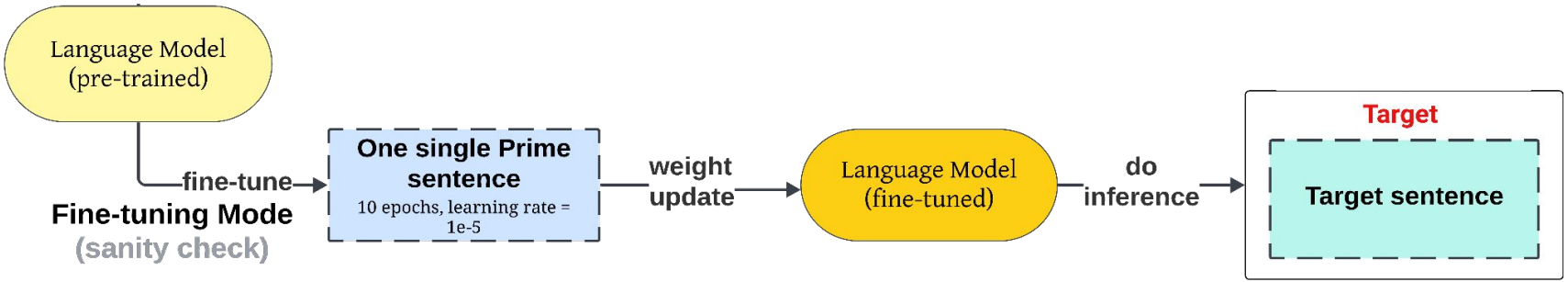
⇒ The IFE as a diagnostic of the error-driven learning mechanism in ICL!

Current Approach: Overview

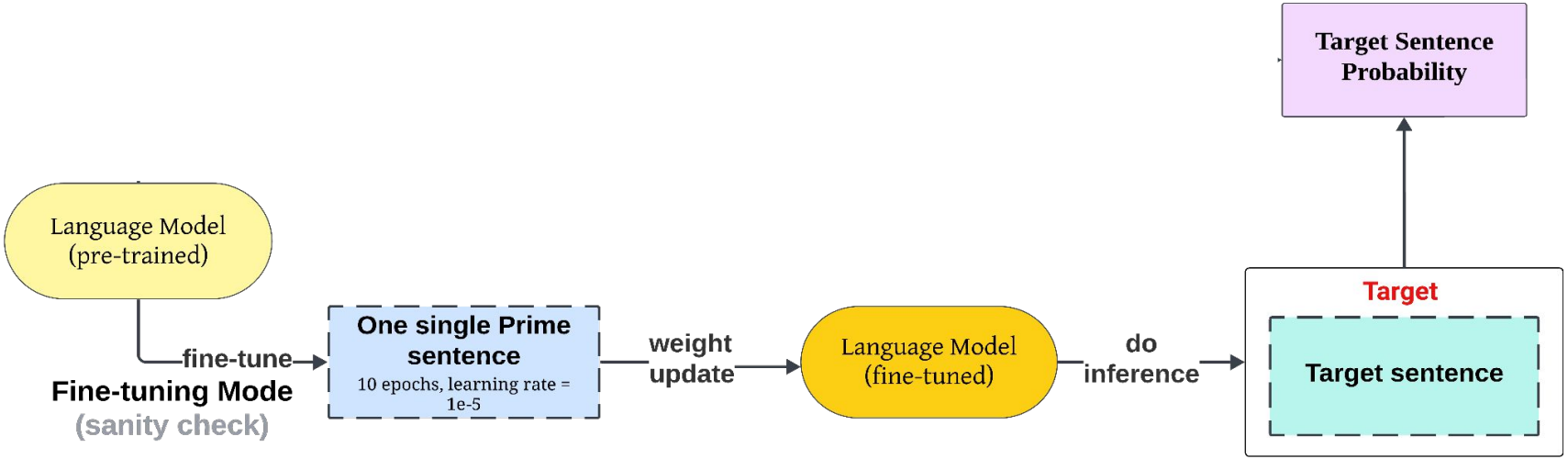
Current Approach: Overview



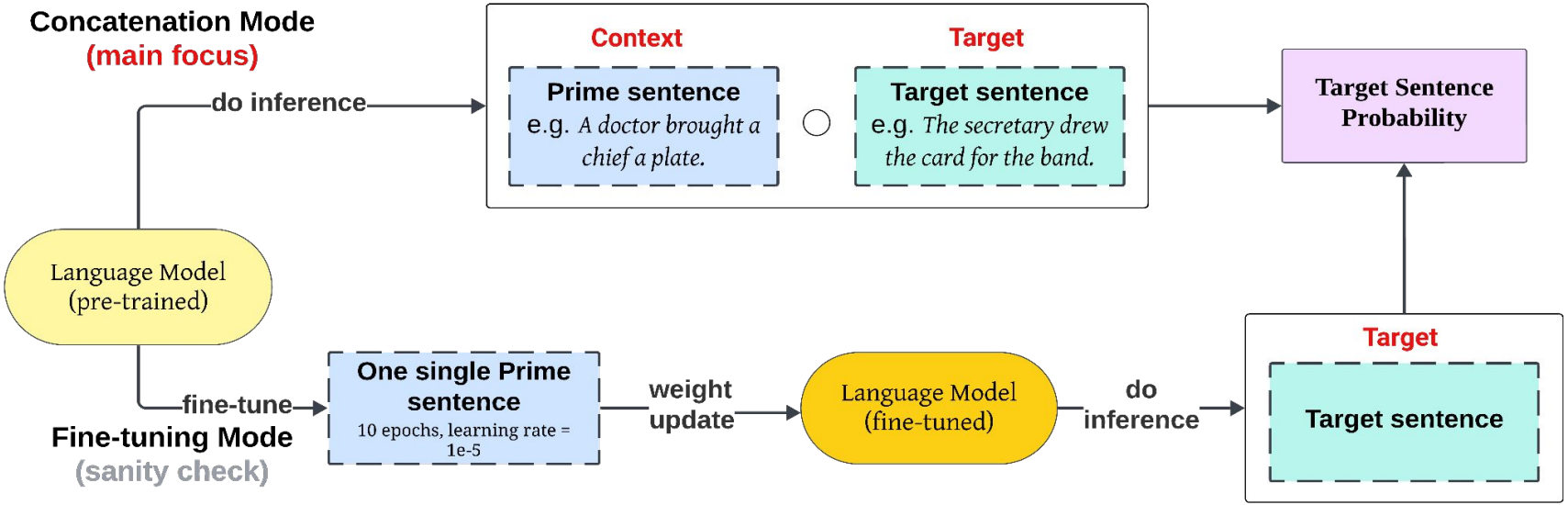
Current Approach: Overview



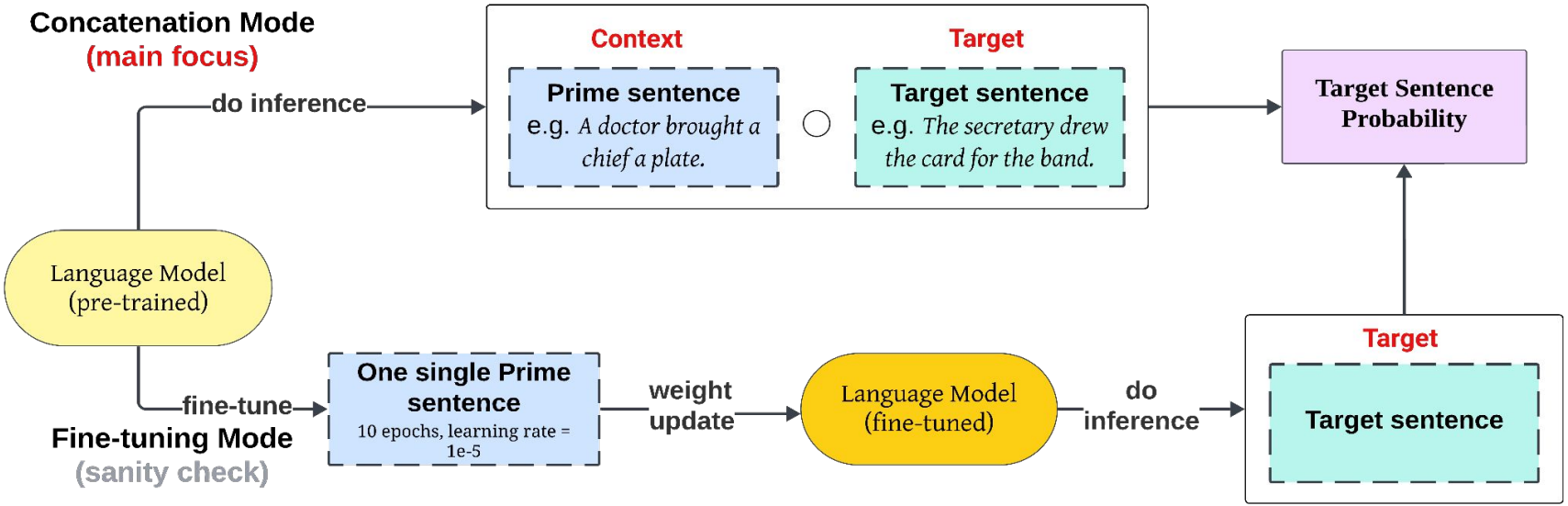
Current Approach: Overview



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

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

Dataset adapted from 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

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

DO prime + DO target	DO prime + PD target
PD prime + DO target	PD prime + PD target

Dataset adapted from 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

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

DO prime + DO target	DO prime + PD target
PD prime + DO target	PD prime + PD target

Prime: *A professor promised a student a letter*. **Target**: *The secretary drew the card for the boss.*

Dataset adapted from Sinclair et al. 2022

Quantifying Verb Biases and the IFE

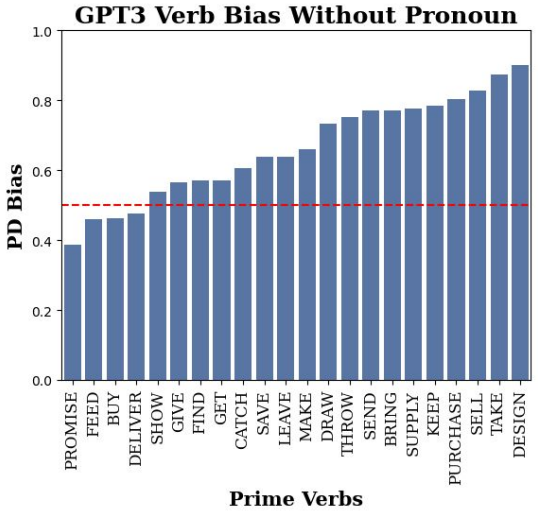
Verb Bias of verb *V* on structure *X*:

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

Quantifying Verb Biases and the IFE

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

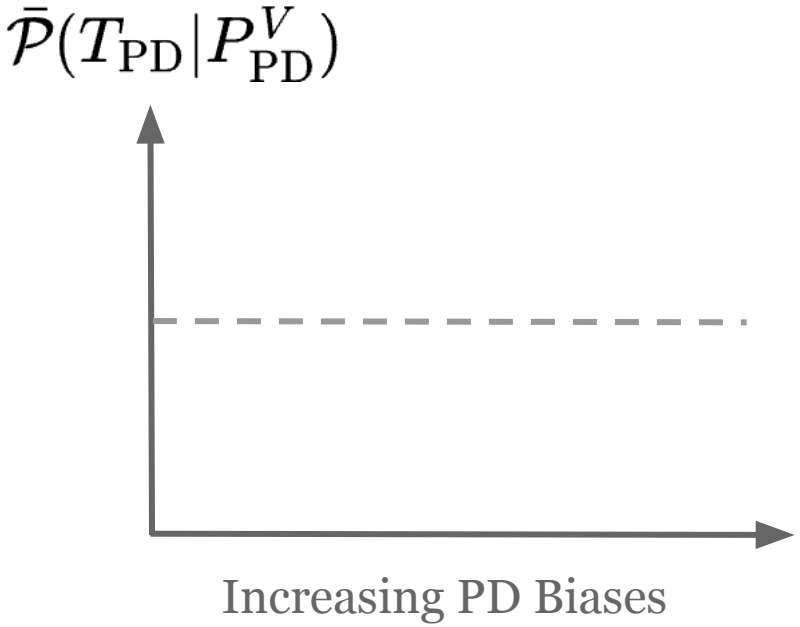
Verb Bias of verb *V* on structure *X*:

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

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

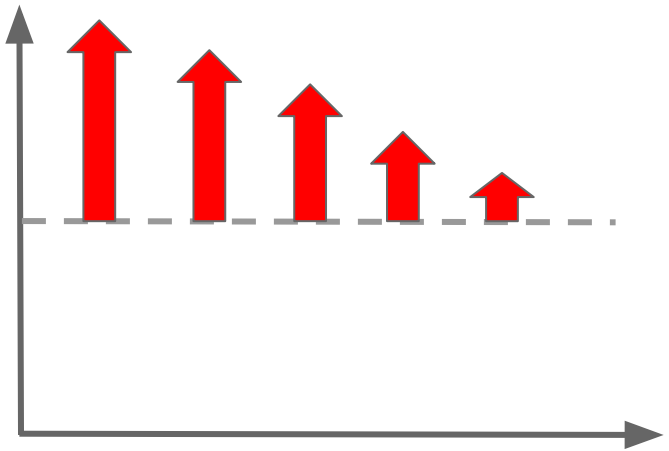
$$PrimeBias(\text{PD}|\text{DO}, V) = \frac{1}{|T_{\text{PD}}| \cdot |P^V_{\text{DO}}|} \sum_{t_{\text{PD}} \in T_{\text{PD}}} \sum_{p^V_{\text{DO}} \in P^V_{\text{DO}}} \frac{\mathcal{P}(t_{\text{PD}}|p^V_{\text{DO}})}{\mathcal{P}(t_{\text{DO}}|p^V_{\text{DO}}) + \mathcal{P}(t_{\text{PD}}|p^V_{\text{DO}})}$$

Predictions on the IFE



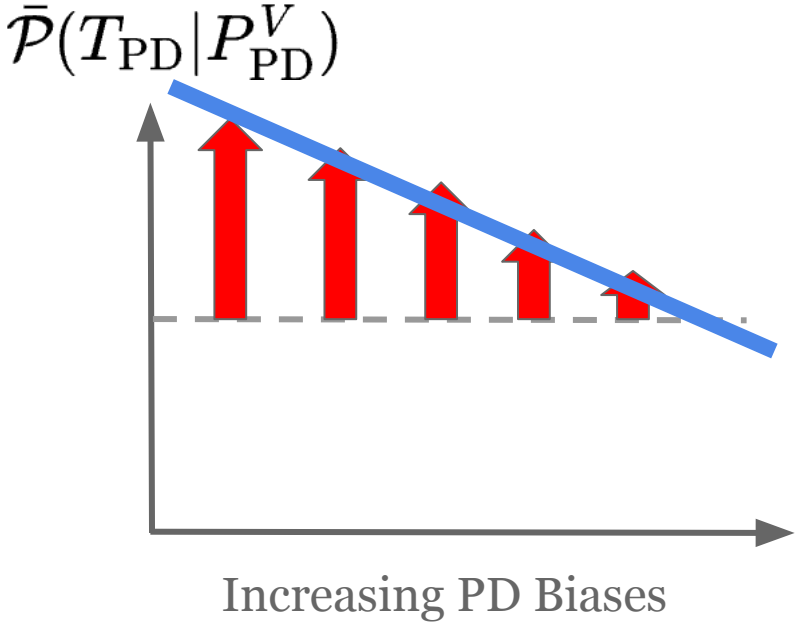
Predictions on the IFE

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

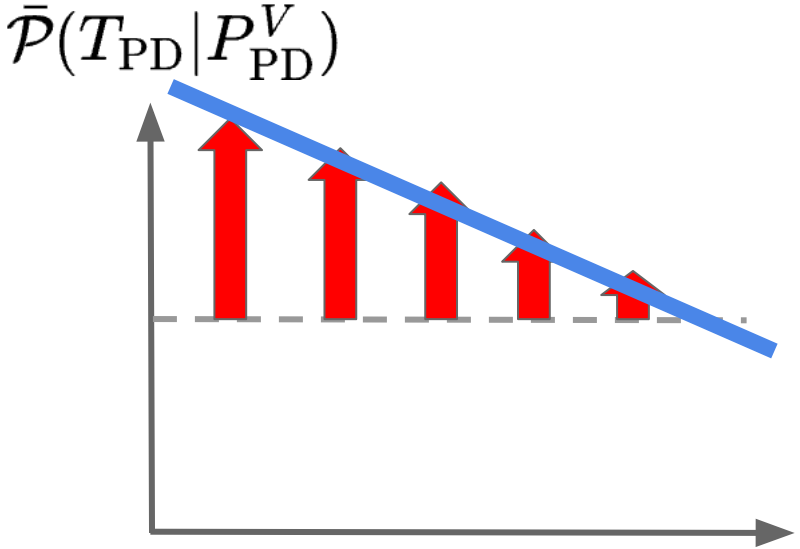


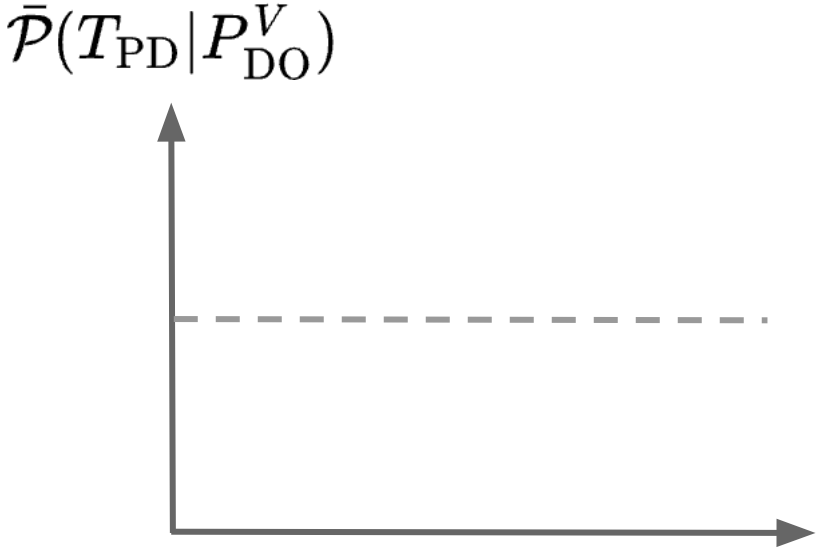
Increasing PD Biases

Predictions on the IFE

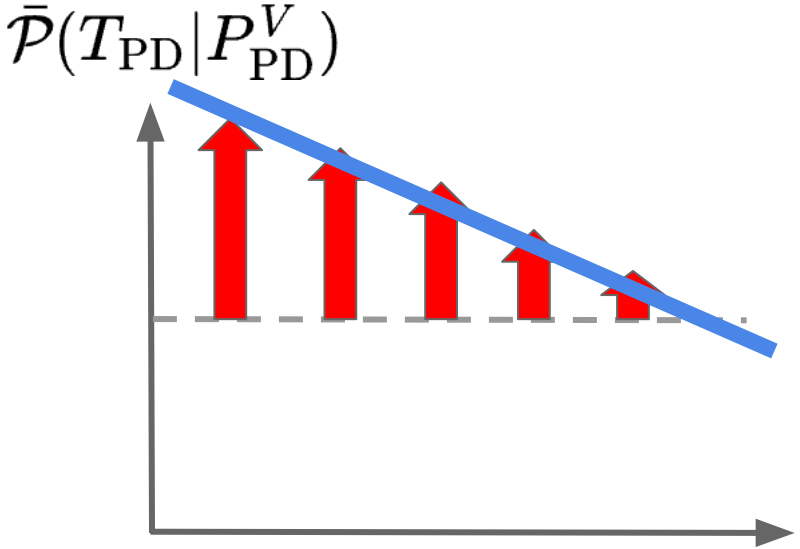


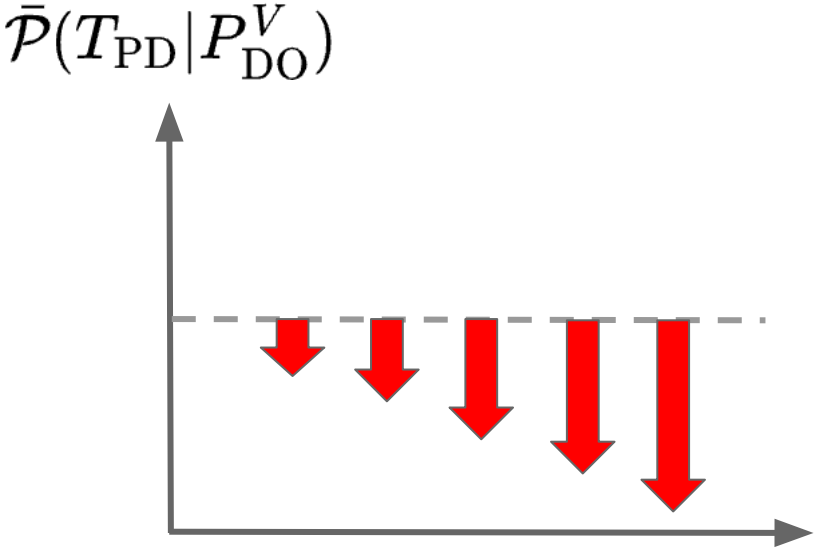
Predictions on the IFE





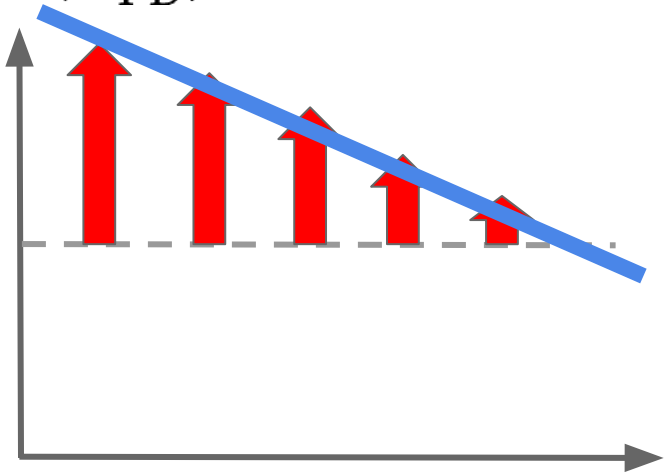
Predictions on the IFE





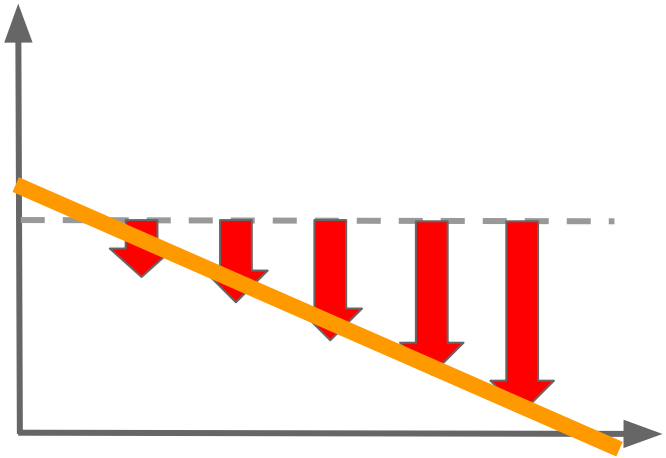
Predictions on the IFE

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



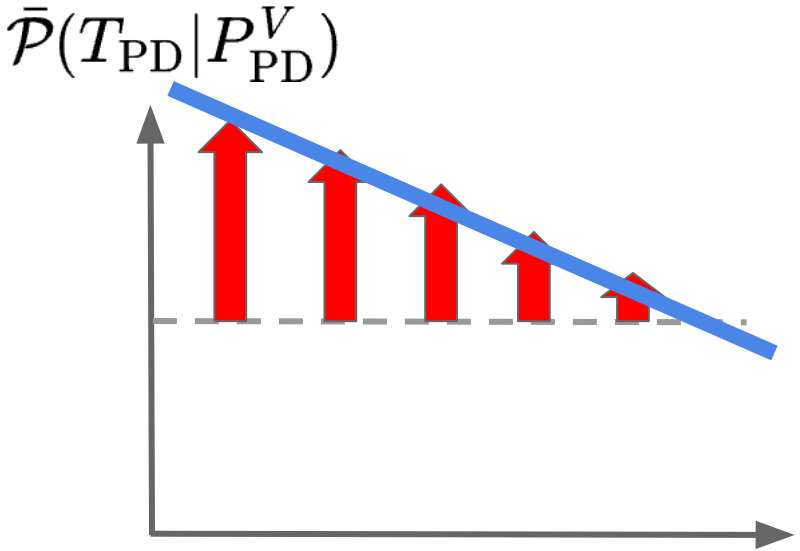
Increasing PD Biases

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

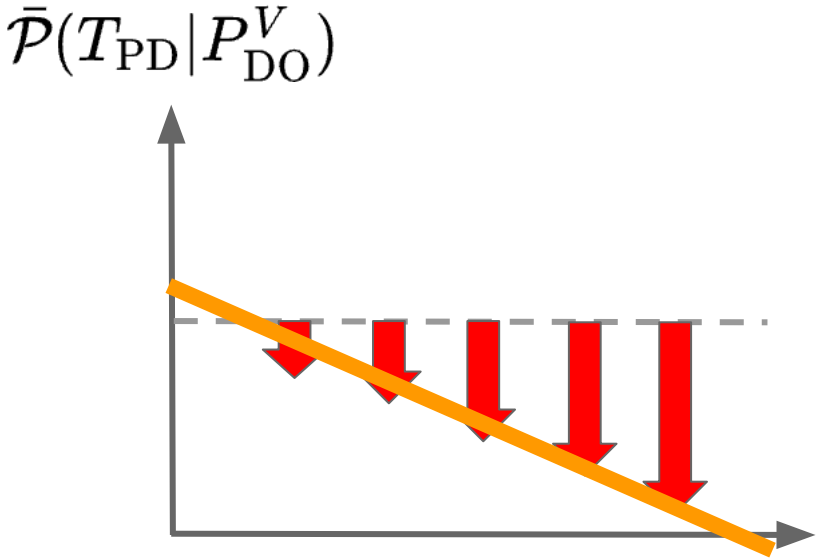


Increasing PD Biases

Predictions on the IFE



Increasing PD Biases



Increasing PD Biases

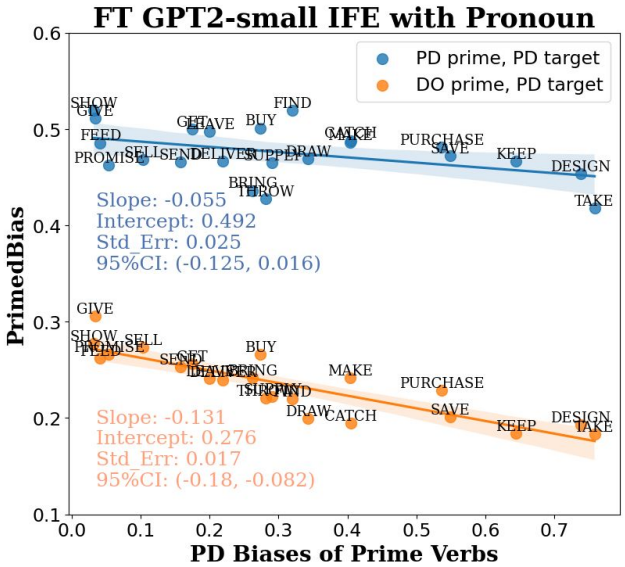
- **IFE:** double negative slopes;
- **Standard Priming:** PD-PD has higher intercept than DO-PD;

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;

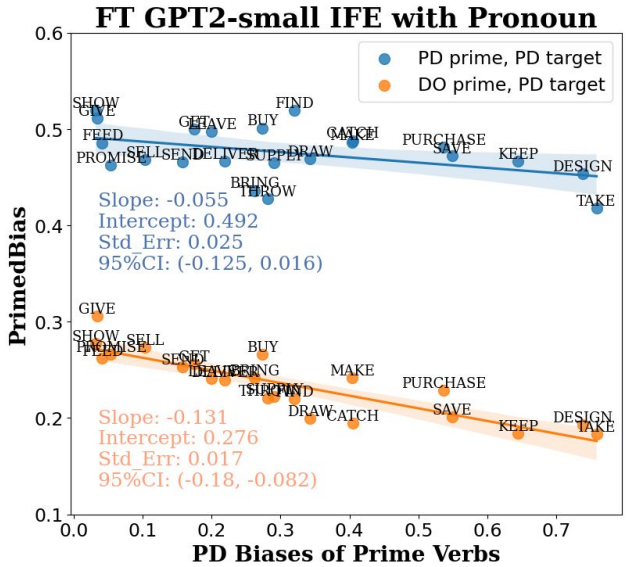
Priming in LLMs

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;



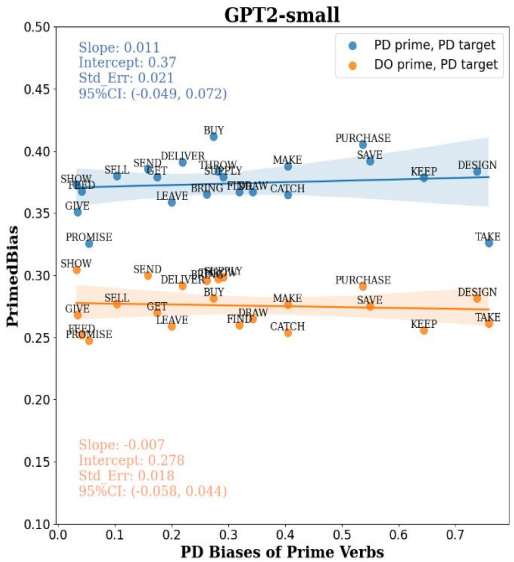
Even *GPT2-small* shows significant inverse frequency effects!

Priming in LLMs

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

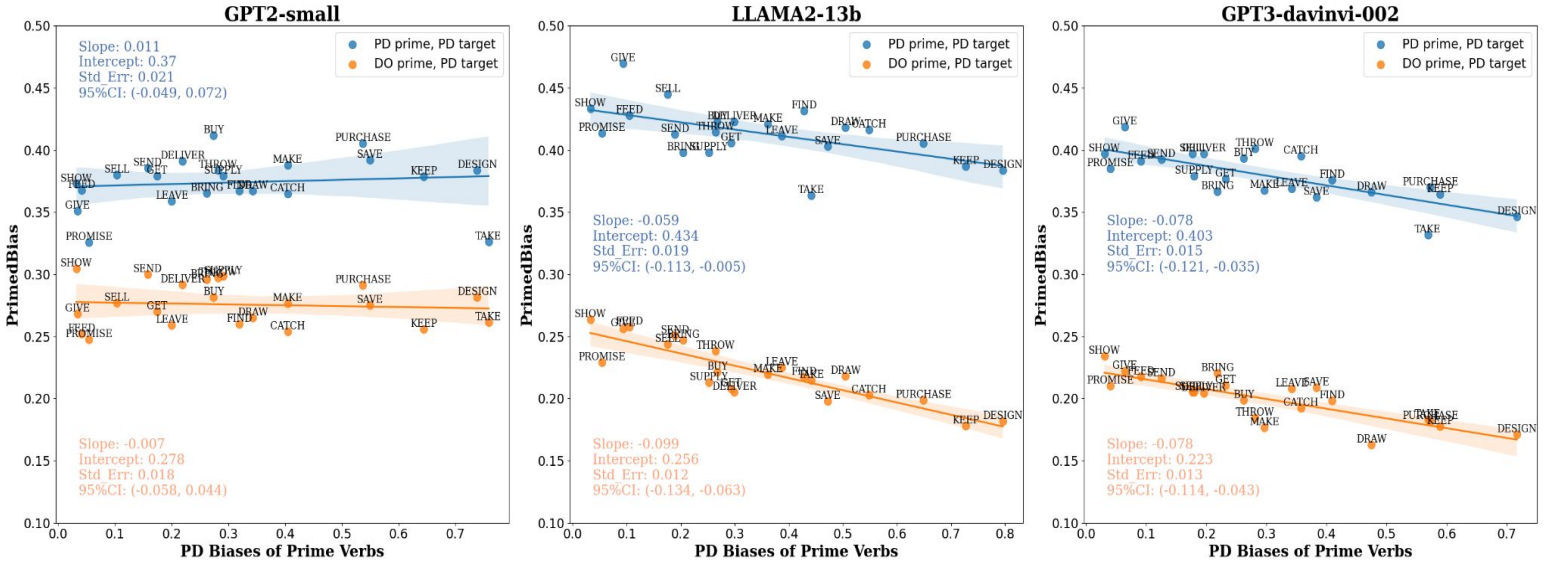
Priming in LLMs

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



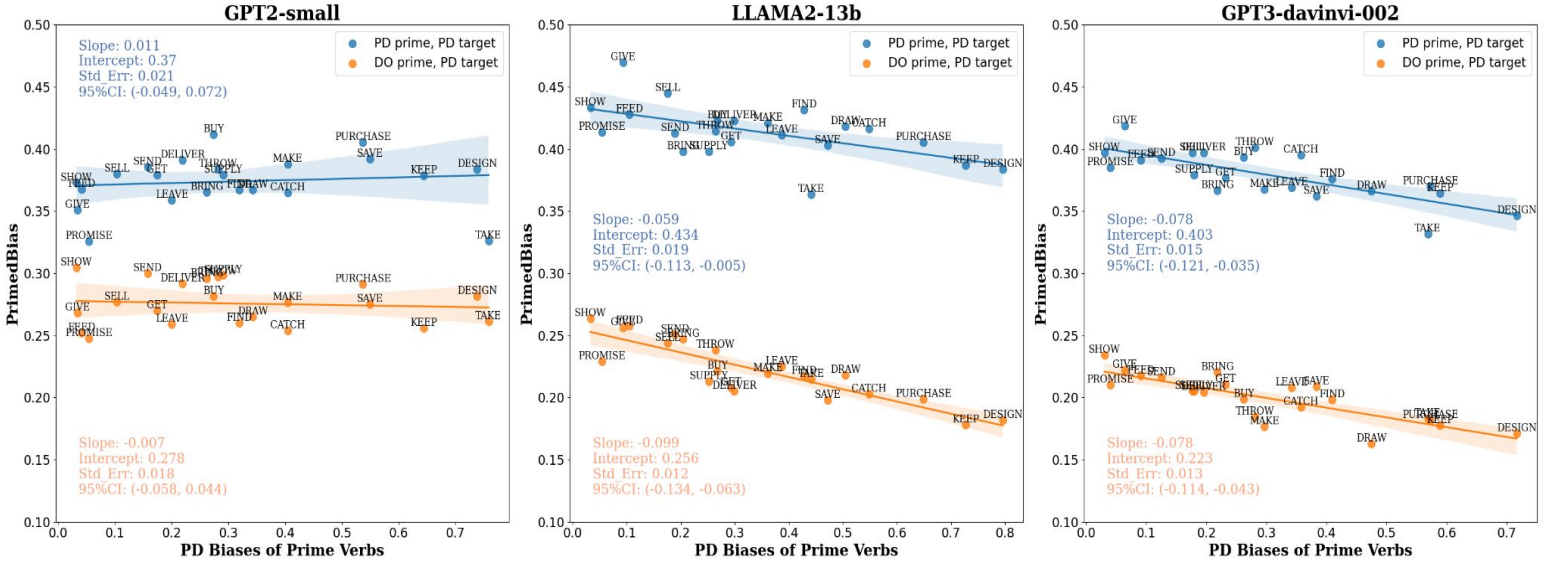
Priming in LLMs

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



Priming in LLMs

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



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

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.

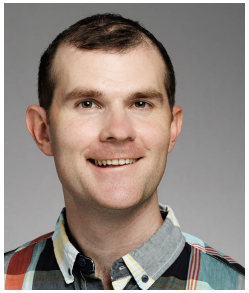
- Generalizing beyond standard notion of ICL, connecting priming with prompting 🌓
- Larger LLMs show *more significant* IFE 🤭🧐

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!🌟🌟🌟

Thanks for Listening!



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;	• Structural priming effect; • The Inverse Frequency Effect (IFE); • Implicit learning, a type of error-driven learning, accounts for the IFE;;	• 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.