ICML
International Conference
On Machine Learning
Vienna, 2024

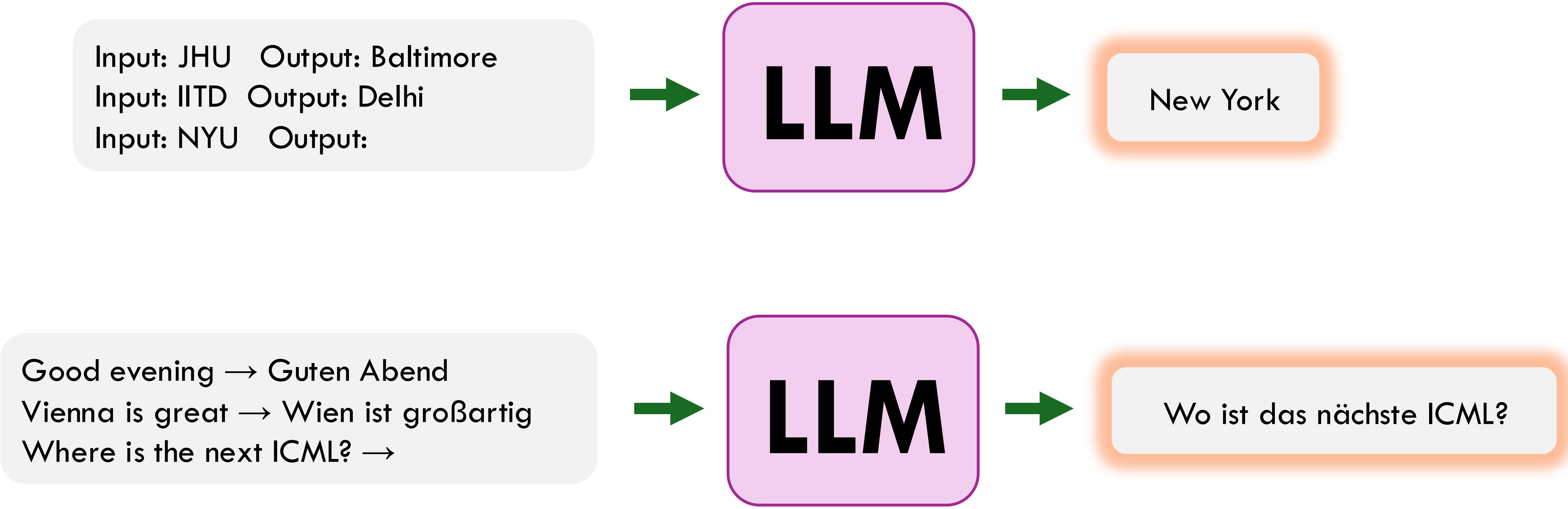
JOHNS HOPKINS
UNIVERSITY

# Position: Do pretrained transformers learn in-context by Gradient Descent?

Lingfeng Shen*, **Aayush Mishra***, Daniel Khashabi

I have a dream that one day … → **LLM** → this nation will rise up …

# In-Context Learning (ICL)



# Transformers Learn Nonlinear Features In Context: Nonconvex Mean-field Dynamics on the Attention Landscape

Juno Kim [1 2] Taiji Suzuki [1 2]

Langua

Damai Dai[†,*]; Y

V

ING? INVESTIGATIONS WITH LINEAR MODELS

Ekin Akyürek[1,2,a] Dale Schuurmans[1] Jacob Andreas[*2] Tengyu Ma[*1,3,b] Denny Zhou[*1]

# Transformers Learn In-Context by Gradient Descent

Johannes von Oswald [1 2] Eyvind Niklasson [2] Ettore Randazzo [2] João Sacramento [1]
Alexander Mordvintsev [2] Andrey Zhmoginov [2] Max Vladymyrov [2]

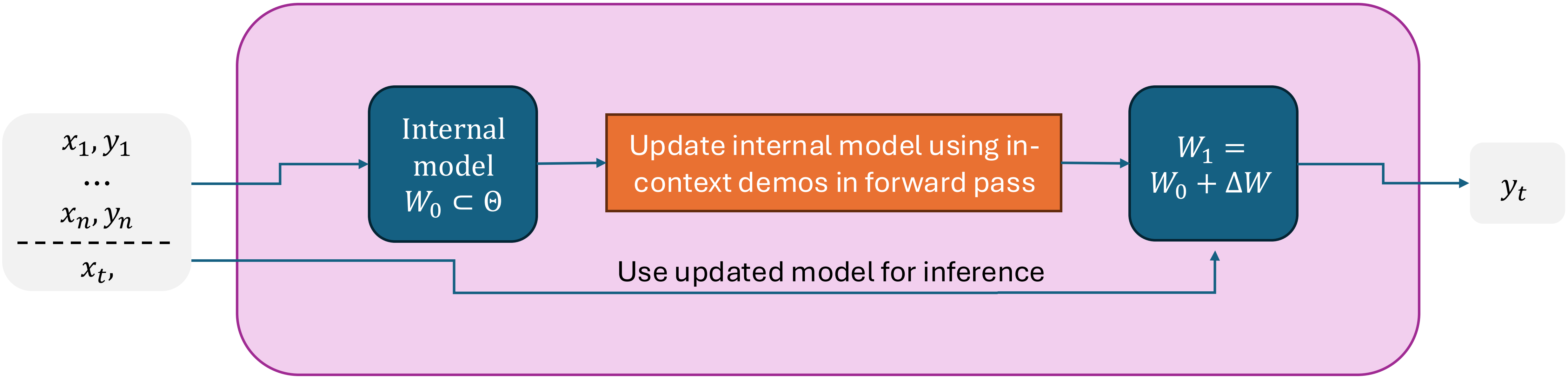
1. Dai, Damai, et al. "Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers." arXiv preprint arXiv:2212.10559 (2022).
2. Akyürek, Ekin, et al. "What learning algorithm is in-context learning? investigations with linear models." arXiv preprint arXiv:2211.15661 (2022).
3. Von Oswald, Johannes, et al. "Transformers learn in-context by gradient descent." International Conference on Machine Learning. PMLR, 2023.
4. Kim, Juno, and Taiji Suzuki. "Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape." arXiv preprint arXiv:2402.01258 (2024).

# The argument

**LLM**

**(weights Θ)**

# The argument



# But here's the thing …

These theories make oversimplified and sometimes unrealistic assumptions.

The functional nature of ICL
(and its equivalence to GD) remains unclear.

# Evolution of this theory

1. *In-context Learning can be* ***interpreted*** *as implicit finetuning*. [Dai+, 2022]

   Show that transformer attention has a dual form of gradient descent:

$$\mathcal{F}(\mathrm{x}) = (\mathrm{W}_0 + \Delta\mathrm{W})\mathrm{x} = \mathrm{W}_0\mathrm{x} + \mathrm{LinearAttn}(\mathrm{E}, \mathrm{X}', \mathrm{x})$$

2. *Hand crafted transformer weights that simulate gradient descent*. [Akyurek+, 2022]

   These weights can **imitate** GD in the forward pass of the transformer.

3. *Weights found by optimization match the construction*. [Oswald+, 2023]

   Compare actual trained weights with their new construction.

   Claim: Gradient-based optimization and attention-based in-context learning are **equivalent**.


1. Dai, Damai, et al. "Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers." arXiv preprint arXiv:2212.10559 (2022).
2. Akyürek, Ekin, et al. "What learning algorithm is in-context learning? investigations with linear models." arXiv preprint arXiv:2211.15661 (2022).
3. Von Oswald, Johannes, et al. "Transformers learn in-context by gradient descent." International Conference on Machine Learning. PMLR, 2023.

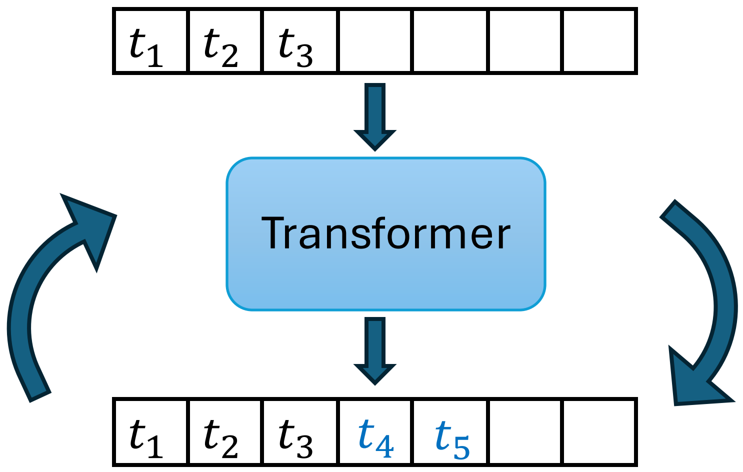
# The “ICL Objective” problem

## ICL

Given an unsupervised corpus of tokens $\{t_1, t_2, \dots t_n\}$, causal language modeling (CLM) objective is used to train the model (with a context window of size $k$):

$$arg\ max_{\Theta} \sum_i \log P(t_i | t_{i-k}, \dots, t_{i-1}; \Theta)$$

- Models trained on unstructured sequences
- Emergent phenomenon




**Transformers, pretrained with CLM objective, yield emergent ICL.**

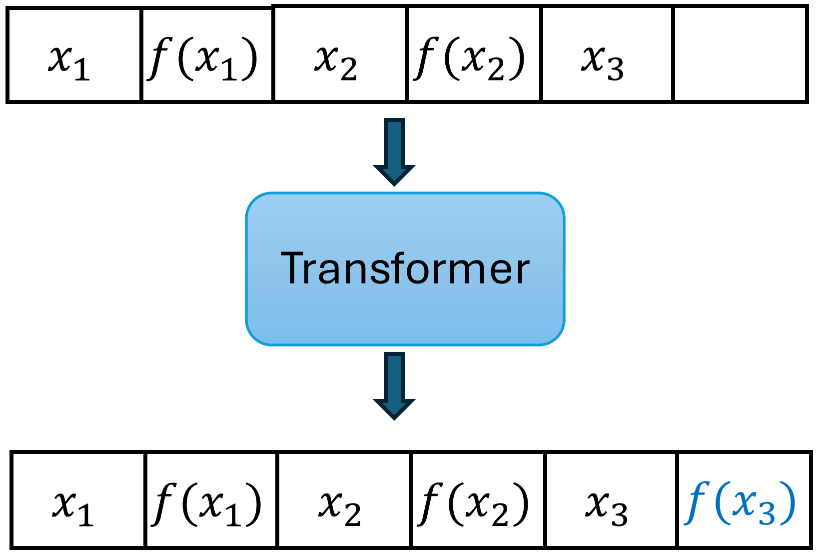
Our new terminology

$\widehat{\text{ICL}}$

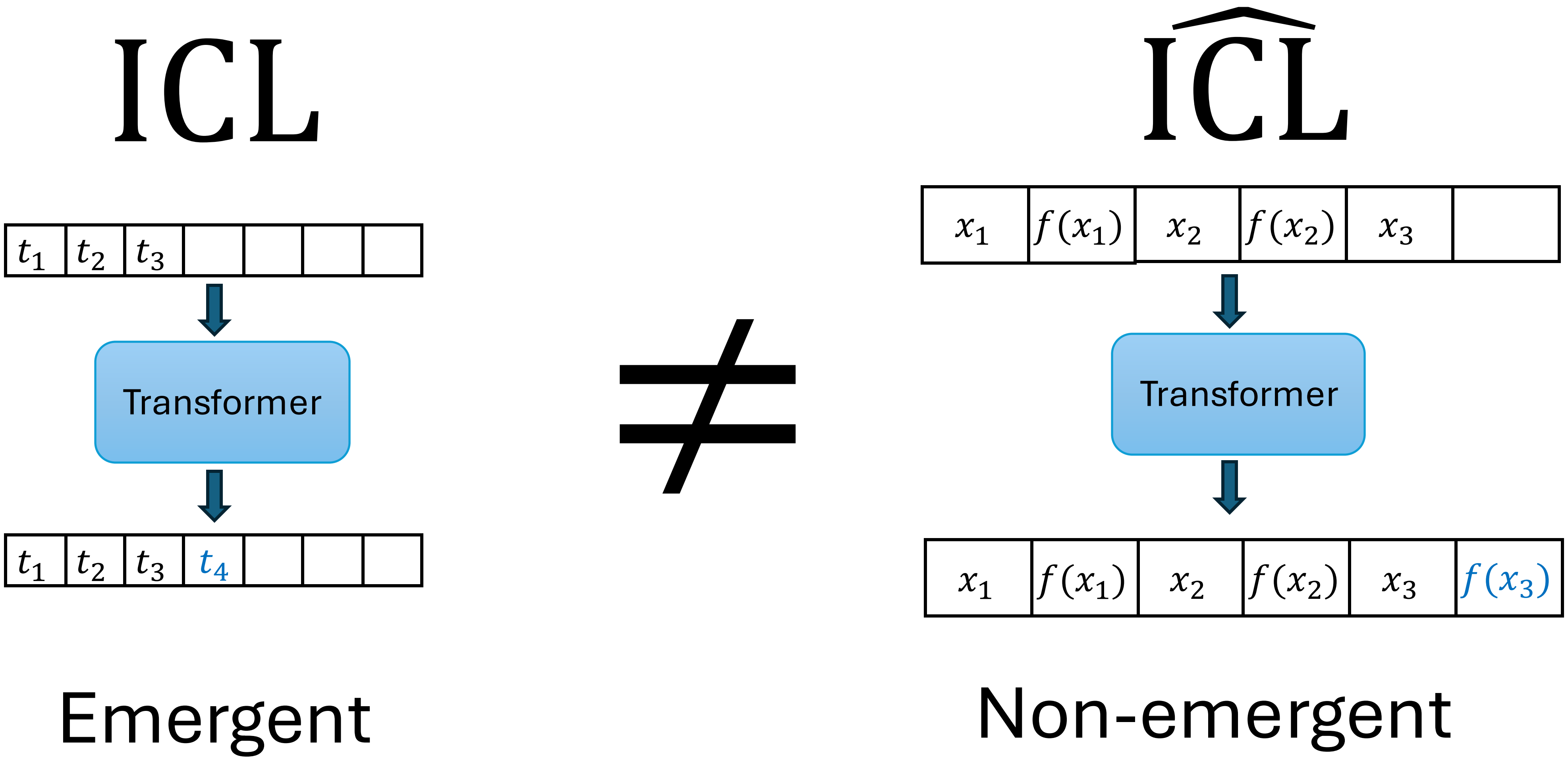
Given an input domain $x \sim X$, and a function class $f \sim F$, ICL objective is used to train the model by giving it structured paired inputs:

$$arg\ max_{\Theta} \sum \log P(f(x_{n+1}) |\ x_1, f(x_1), \dots, x_n \circ f(x_n) \circ x_{n+1}; \Theta)$$

- Models trained on structured sequences
- Non-Emergent phenomenon




**Transformers trained with ICL objective, yield non-emergent meta learning.**



ICL
$t_1$ $t_2$ $t_3$
Transformer
$t_1$ $t_2$ $t_3$ $t_4$
Emergent
$\neq$
$\widehat{\text{ICL}}$
$x_1$ $f(x_1)$ $x_2$ $f(x_2)$ $x_3$
Transformer
$x_1$ $f(x_1)$ $x_2$ $f(x_2)$ $x_3$ $f(x_3)$
Non-emergent

What they imply

## $\mathrm{ICL} \approx \mathrm{GD}$ equivalence

**For any** Transformer weights resulting from self-supervised pretraining and **for any** well-defined task, $\mathrm{ICL}$ is algorithmically equivalent to $\mathrm{GD}$.

What they show

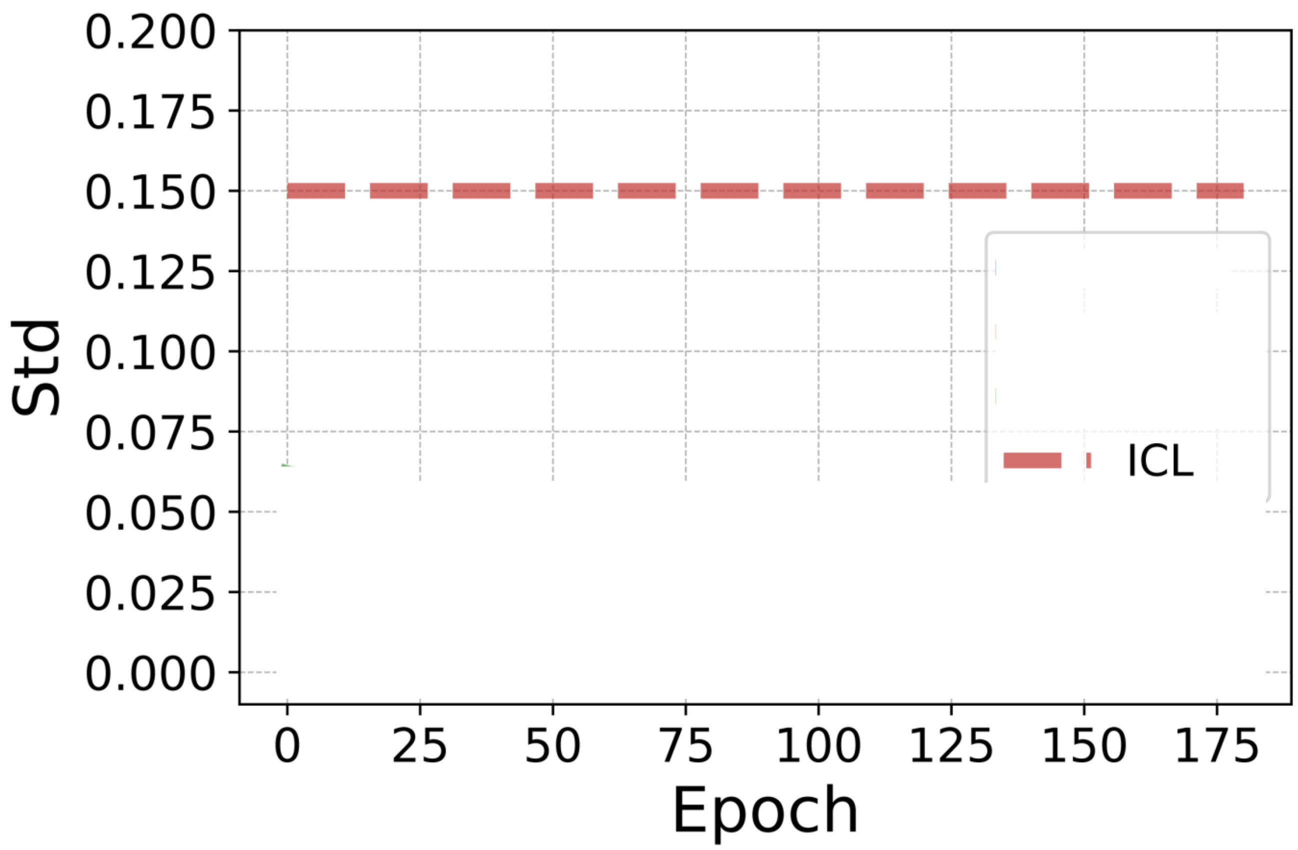
## $\widehat{\mathrm{ICL}} \approx \mathrm{GD}$ equivalence

For a **given** well-defined task, **there exist** Transformer weights such that $\widehat{\mathrm{ICL}}$ is algorithmically equivalent to $\mathrm{GD}$.

Transformers have the expressive capacity to simulate GD.

This does not imply that LLMs actually do simulate it.

# Evidence against $\mathrm{ICL} \approx \mathrm{GD}$: Order sensitivity




- ICL is known to be highly order-sensitive [Lu+].
- GD is order-insensitive.
  contradicts [Oswald+]
- Variants of GD are still not as sensitive as ICL.
  undermines [Akyurek+]

ICL is likely **not** equivalent to GD based on order sensitivity.



2. Akyürek, Ekin, et al. "What learning algorithm is in-context learning? investigations with linear models." arXiv preprint arXiv:2211.15661 (2022).
3. Von Oswald, Johannes, et al. "Transformers learn in-context by gradient descent." International Conference on Machine Learning. PMLR, 2023.
4. Lu, Yao, et al. "Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity." arXiv preprint arXiv:2104.08786 (2021).

# Evidence against ICL ≈ GD: Weight Sparsity

- Hand constructed weights and inputs in [2] are **highly sparse.**
- Constructions of [3] are similarly sparse.

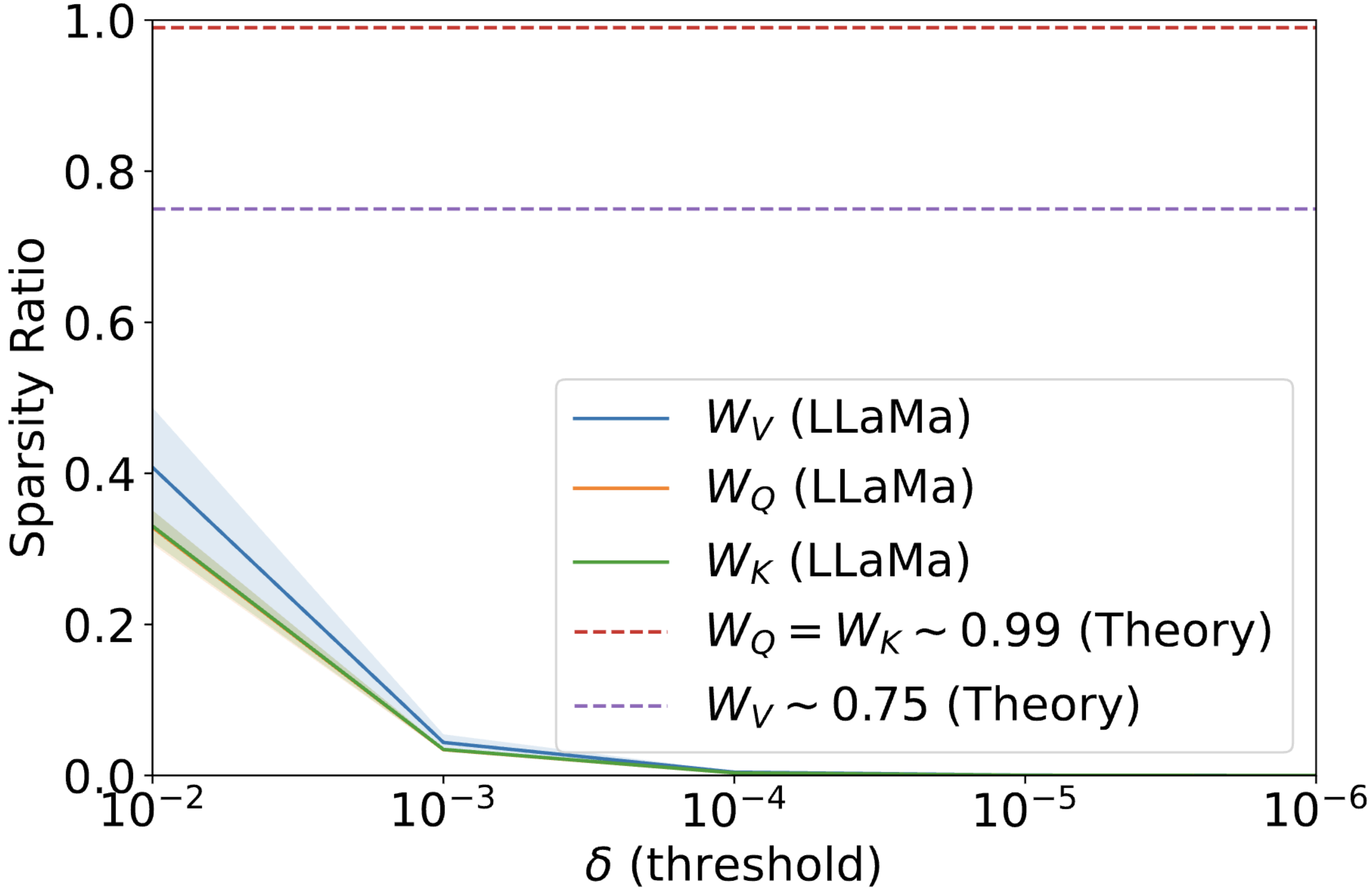
$$H^{(0)} = \begin{bmatrix} \cdots & \boxed{0} & y_i & \boxed{0} & \cdots \\ & \boldsymbol{x}_i & \boxed{0} & \boldsymbol{x}_n & \end{bmatrix} \qquad W_e = \begin{pmatrix} \boxed{I^{(d+1)\times(d+1)}} & \boxed{\mathbf{0}} \\ \boxed{\mathbf{0}} & \boxed{\mathbf{0}} \end{pmatrix} \qquad W_V = \begin{pmatrix} \boxed{0} & \boxed{0} \\ W_0 & \boxed{-I_y} \end{pmatrix}$$

$$W_K^l = \begin{pmatrix} \mathbf{0} & \cdots & & \\ & \ddots & & \\ & I^{p\times p} & \mathbf{0}^{p\times p} & \\ & & & \cdots \\ & & \vdots & \end{pmatrix} \qquad W_Q^l = \begin{pmatrix} \mathbf{0} & \cdots & & \\ & \ddots & & \\ & \mathbf{0}^{p\times p} & I^{p\times p} & \\ & & & \cdots \\ & & \vdots & \end{pmatrix} \qquad W_K = W_Q = \begin{pmatrix} \boxed{I_x} & \boxed{0} \\ \boxed{0} & \boxed{0} \end{pmatrix}$$

$$P = \frac{\eta}{N} \boxed{I}$$

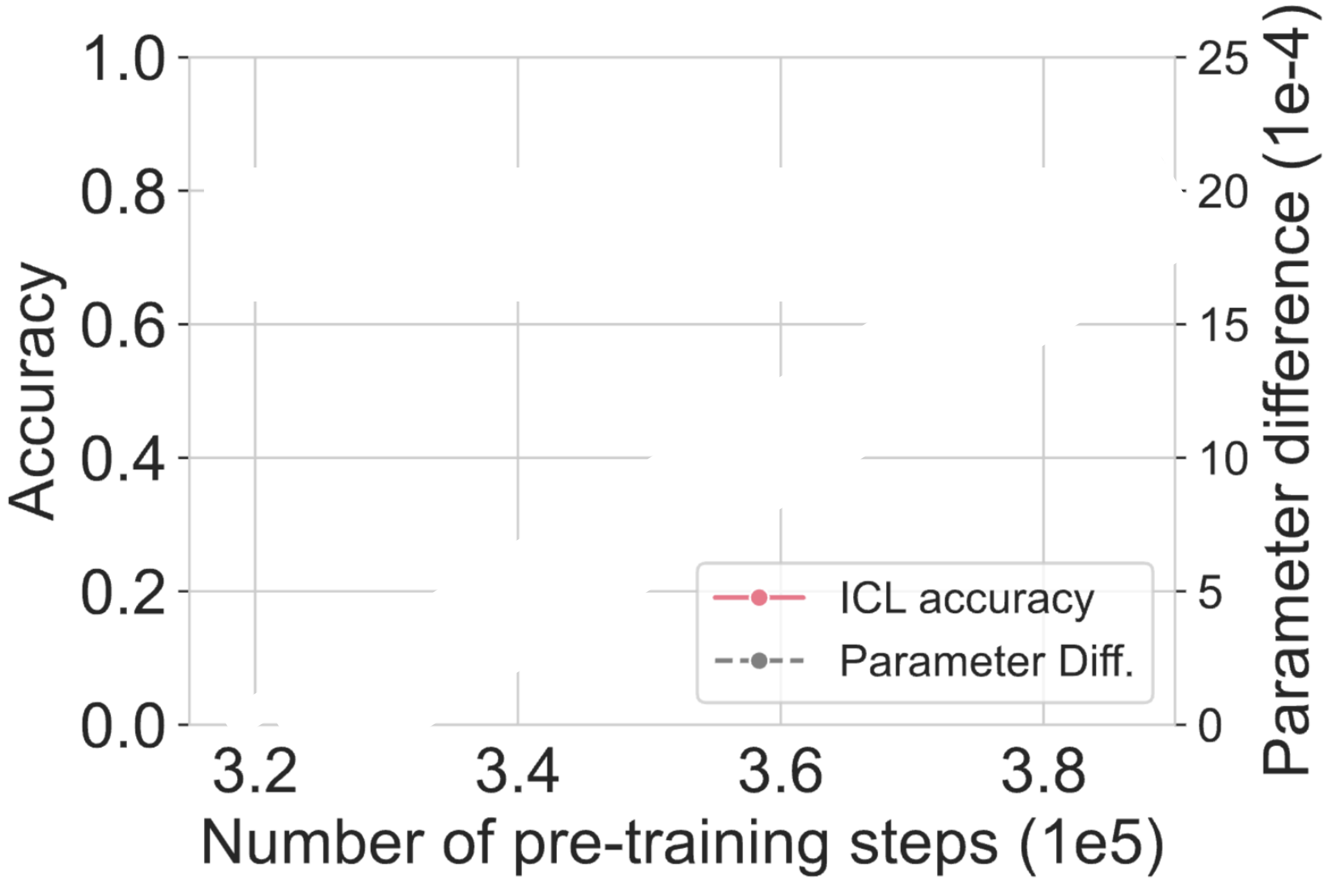

2. Akyürek, Ekin, et al. "What learning algorithm is in-context learning? investigations with linear models." arXiv preprint arXiv:2211.15661 (2022).
3. Von Oswald, Johannes, et al. "Transformers learn in-context by gradient descent." International Conference on Machine Learning. PMLR, 2023.

# Evidence against $\mathrm{ICL} \approx \mathrm{GD}$: Weight Sparsity




Real LLMs are rather dense.

# Evidence against $\mathrm{ICL} \approx \mathrm{GD}$: Weight Evolution




ICL ability remains stable even when weights keep evolving.

To claim $\mathrm{ICL} \approx \mathrm{GD}$, showing it for a single sparse choice of parameters is **not** enough.

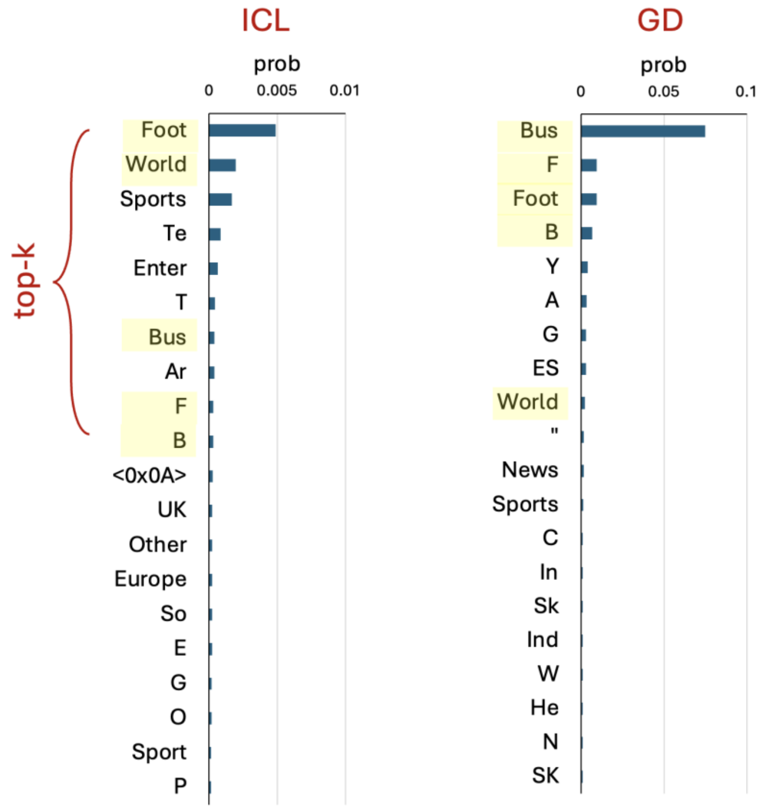
# Evidence against $\mathrm{ICL} \approx \mathrm{GD}$: Performance

We use three coarse-to-fine metrics to compare ICL and GD:

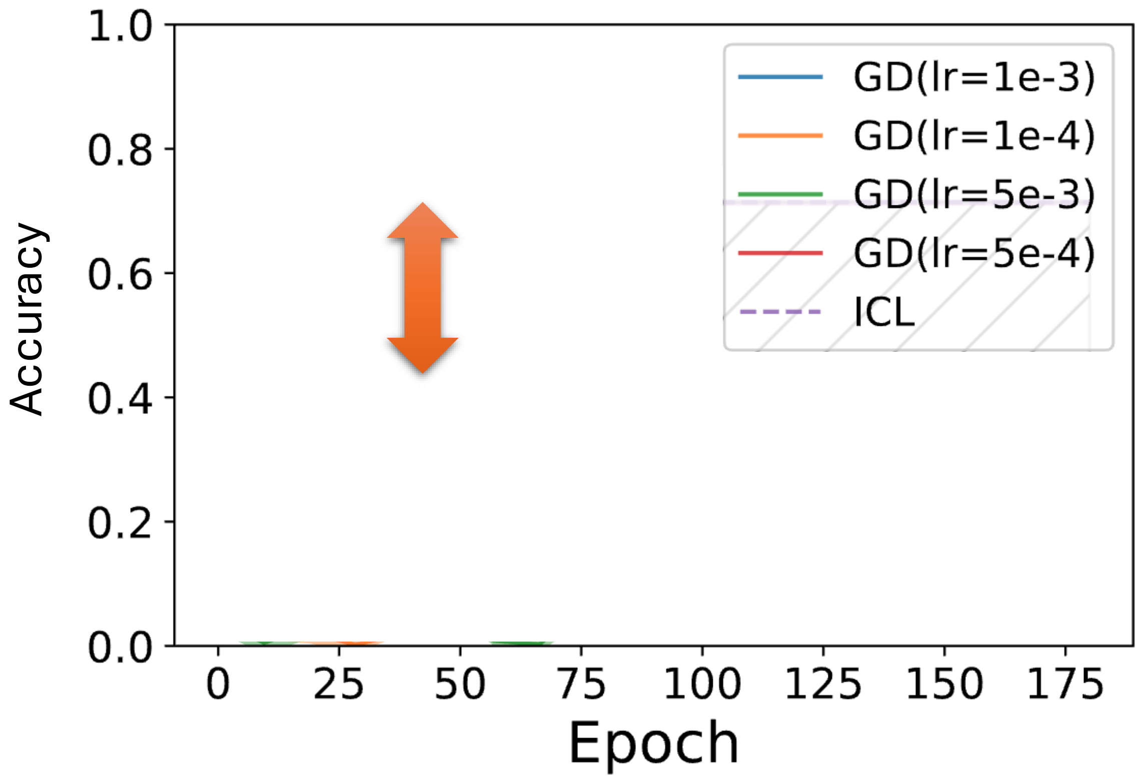
1. **Accuracy**: compare top predicted tokens against the ground truth (instead of top predicted label).

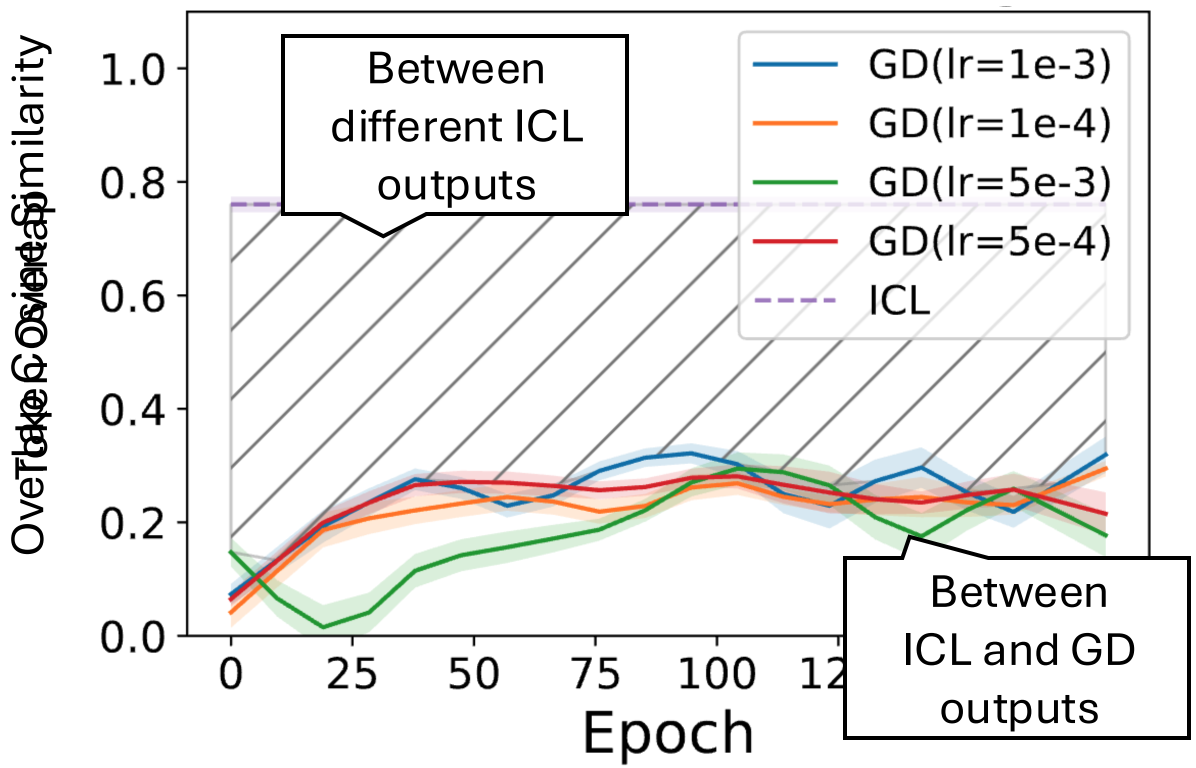
2. **Token Overlap**: compare overlap in top-K tokens of two distributions.

3. **Overlap Cosine Similarity**: compare the individual agreement between top-K tokens.



# Evidence against $\mathrm{ICL} \approx \mathrm{GD}$: Outputs




ICL performs differently and does not align with GD.

# Summary

1. Recent works studying ICL **do not align with emergent ICL in LLMs.**
   In-Context Learning → Learning to Learn, Meta Learning, etc.
   Transformers learn in-context by gradient descent → Transformers can perform gradient descent in their forward pass when trained appropriately.

2. Expressivity of the Transformer architecture to simulate GD **does not imply** that LLMs actually do it.
   We present arguments and evidence against the [current] $\mathrm{ICL} \approx \mathrm{GD}$ equivalence theory.

3. Maintain **parallels to real world** settings when developing.
   It is OK to study ICL in a simpler setting like Linear Regression,
   but need to find a corresponding pretraining distribution to elicit ICL.




2. Akyürek, Ekin, et al. "What learning algorithm is in-context learning? investigations with linear models." arXiv preprint arXiv:2211.15661 (2022).
3. Von Oswald, Johannes, et al. "Transformers learn in-context by gradient descent." International Conference on Machine Learning. PMLR, 2023.

# Thank you!

Paper

Contact