State-of-the-art Seminar

Understanding In-Context Learning

From the Architecture Perspective

Saqib Sarwar

August 16, 2025

Outline

Table of Contents

In-Context Learning

Motivation Example

Answer in One Word.
"You have a right to perform your prescribed duties, but you are not entitled to the fruits of your actions." : Bhagavad Gita
"Love is patient, love is kind. It does not envy, it does not boast, it is not proud." : Corinthians
"The root of suffering is attachment." : Samyutta Nikaya
"And your Lord never forgets." : ?

Qur'an

In-Context Learning

Motivation Example

Answer in One Word.
"You have a right to perform your prescribed duties, but you are not entitled to the fruits of your actions." : Bhagavad Gita
"Love is patient, love is kind. It does not envy, it does not boast, it is not proud." : Corinthians
"The root of suffering is attachment." : Samyutta Nikaya
"And your Lord never forgets." : ?

Qur'an

"Love is patient, love is kind. It does not envy, it does not boast, it is not proud." : Christianity
"The root of suffering is attachment." : Buddhism
"And your Lord never forgets." : Islam
"You have a right to perform your prescribed duties, but you are not entitled to the fruits of your actions." : ?

Hinduism

In-Context Learning

Motivation Example

Answer in One Word.
"You have a right to perform your prescribed duties, but you are not entitled to the fruits of your actions." : Bhagavad Gita
"Love is patient, love is kind. It does not envy, it does not boast, it is not proud." : Corinthians
"The root of suffering is attachment." : Samyutta Nikaya
"And your Lord never forgets." : ?

Qur'an

"Love is patient, love is kind. It does not envy, it does not boast, it is not proud." : Christianity
"The root of suffering is attachment." : Buddhism
"And your Lord never forgets." : Islam
"You have a right to perform your prescribed duties, but you are not entitled to the fruits of your actions." : ?

Hinduism

"You have a right to perform your prescribed duties, but you are not entitled to the fruits of your actions." : Detachment
"And your Lord never forgets." : Omniscient
"Love is patient, love is kind. It does not envy, it does not boast, it is not proud." : Benevolent
"The root of suffering is attachment." : ?

Clinging

The Pre-ICL Paradigm

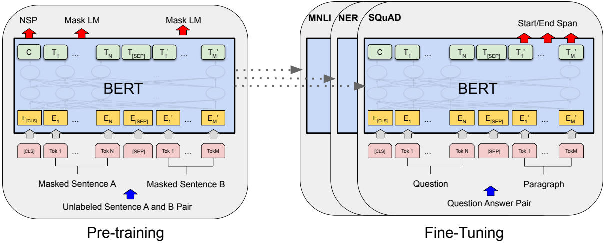


Figure: Pre-training and Supervised Fine-Tuning

1

[1] Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

In-Context Learning

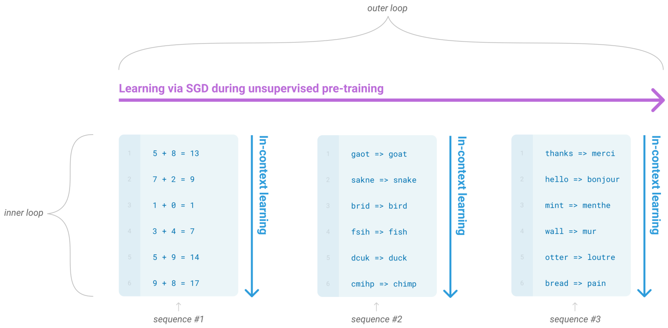


Figure: Language Model Meta Learning

2

[2] Brown, T., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*,

In-Context Learning: Scaling Effects

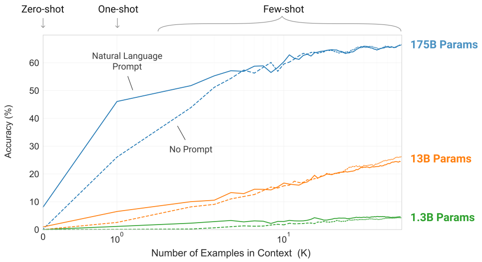


Figure: Word Scrambling and Manipulation Tasks

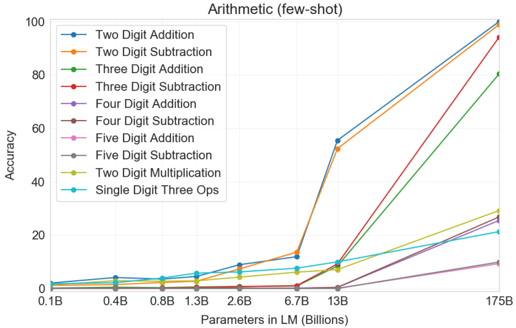


Figure: Arithmetic Tasks

3

[3] Brown, T., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, **33**, 1877–1901.

Hypothesis 1: ICL as a Meta Optimizer

Implicit Fine Tuning

Meta-ICL

	Meta-training	Inference
Task	C *meta-training* tasks	An unseen *target* task
Data given	Training examples $\mathcal{T}_i = \{(x_j^i, y_j^i)\}_{j=1}^{N_i}, \ \forall i \in [1, C] \quad (N_i \gg k)$	Training examples $(x_1, y_1), \cdots, (x_k, y_k)$, Test input x
Objective	For each iteration, 1. Sample task $i \in [1, C]$ 2. Sample $k+1$ examples from $\mathcal{T}_i$: $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ 3. Maximize $P(y_{k+1} \vert x_1, y_1, \cdots, x_k, y_k, x_{k+1})$	$\text{argmax}_{c \in \mathcal{C}} P(c \vert x_1, y_1, \cdots, x_k, y_k, x)$

Figure: Meta-ICL Task

[4] Min, Sewon, et al. "MetaICL: Learning to Learn In Context." *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 2791–2809.

Meta-ICL

	Meta-training	Inference
Task	C *meta-training* tasks	An unseen *target* task
Data given	Training examples $\mathcal{T}_i = \{(x_j^i, y_j^i)\}_{j=1}^{N_i}, \ \forall i \in [1, C] \quad (N_i \gg k)$	Training examples $(x_1, y_1), \cdots, (x_k, y_k)$, Test input x
Objective	For each iteration, 1. Sample task $i \in [1, C]$ 2. Sample $k+1$ examples from $\mathcal{T}_i$: $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ 3. Maximize $P(y_{k+1} \vert x_1, y_1, \cdots, x_k, y_k, x_{k+1})$	$\text{argmax}_{c \in \mathcal{C}} P(c \vert x_1, y_1, \cdots, x_k, y_k, x)$

Figure: Meta-ICL Task

Meta-train			Target	
Setting	# tasks	# examples	Setting	# tasks
HR	61	819,200	LR	26
Classification Non-Classification	43 37	384,022 368,768	Classification	20
QA Non-QA	37 33	486,143 521,342	QA	22
Non-NLI	55	463,579	NLI	8
Non-Paraphrase	59	496,106	Paraphrase	4

Figure: Meta-ICL Experiments

4

[4] Min, Sewon, et al. "MetaICL: Learning to Learn In Context." *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 2791–2809.

ICL as Meta Optimizer

- **Algorithmic Equivalence:** Transformers can simulate linear learners (GD, ridge, least-squares), transitioning to Bayesian estimators with depth/width. [5]

[5] Akyürek, Ege, et al. "What Learning Algorithm Is In-Context Learning? Investigations with Linear Models." *The Eleventh International Conference on Learning Representations*, 2023.

[6] Dai, Damai, et al. "Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers." *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[7] Garg, Shivam, et al. "What Can Transformers Learn In-Context? A Case Study of Simple Function Classes." *Advances in Neural Information Processing Systems* 35 (NeurIPS 2022).

ICL as Meta Optimizer

- **Algorithmic Equivalence:** Transformers can simulate linear learners (GD, ridge, least-squares), transitioning to Bayesian estimators with depth/width. [5]
- **Implicit Finetuning:** Transformer attention is dual to gradient descent. [6]
  $\Rightarrow$ ICL $\approx$ internal meta-gradient updates.

---

[5] Akyürek, Ege, et al. "What Learning Algorithm Is In-Context Learning? Investigations with Linear Models." *The Eleventh International Conference on Learning Representations*, 2023.

[6] Dai, Damai, et al. "Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers." *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[7] Garg, Shivam, et al. "What Can Transformers Learn In-Context? A Case Study of Simple Function Classes." *Advances in Neural Information Processing Systems* 35 (NeurIPS 2022).

ICL as Meta Optimizer

- **Algorithmic Equivalence:** Transformers can simulate linear learners (GD, ridge, least-squares), transitioning to Bayesian estimators with depth/width. [5]
- **Implicit Finetuning:** Transformer attention is dual to gradient descent. [6]
  $\Rightarrow$ ICL $\approx$ internal meta-gradient updates.
- **Function Class Generalization:** Transformers trained from scratch can in-context learn full function classes (linear, sparse linear, 2-layer MLPs, decision trees), even under distribution shift. [7]

---

[5] Akyürek, Ege, et al. ”What Learning Algorithm Is In-Context Learning? Investigations with Linear Models.” *The Eleventh International Conference on Learning Representations*, 2023.

[6] Dai, Damai, et al. “Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers.” *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[7] Garg, Shivam, et al. “What Can Transformers Learn In-Context? A Case Study of Simple Function Classes.” *Advances in Neural Information Processing Systems* 35 (NeurIPS 2022).

ICL as Meta Optimizer

- **Algorithmic Equivalence:** Transformers can simulate linear learners (GD, ridge, least-squares), transitioning to Bayesian estimators with depth/width. [5]
- **Implicit Finetuning:** Transformer attention is dual to gradient descent. [6]
  $\Rightarrow$ ICL $\approx$ internal meta-gradient updates.
- **Function Class Generalization:** Transformers trained from scratch can in-context learn full function classes (linear, sparse linear, 2-layer MLPs, decision trees), even under distribution shift. [7]

---

[5] Akyürek, Ege, et al. "What Learning Algorithm Is In-Context Learning? Investigations with Linear Models." *The Eleventh International Conference on Learning Representations*, 2023.

[6] Dai, Damai, et al. "Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers." *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[7] Garg, Shivam, et al. "What Can Transformers Learn In-Context? A Case Study of Simple Function Classes." *Advances in Neural Information Processing Systems* 35 (NeurIPS 2022).

ICL as Meta Optimizer

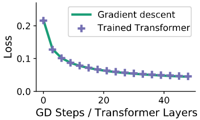


Figure: SGD-Transformer Equivalence

[8] von Oswald, Johannes, et al. "Transformers Learn In-Context by Gradient Descent." *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[9] Wu, Shiguang, et al. "Why In-Context Learning Models are Good Few-Shot Learners?" *International Conference on Learning Representations*, 2025.

ICL as Meta Optimizer

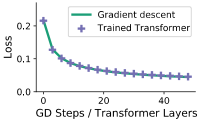


Figure: SGD-Transformer Equivalence

Meta-Learning View: ICL acts as data-dependent meta-learning, distinct from gradient-/metric-/amortized meta-learners. $\Rightarrow$ Implicit algorithm is shaped by pretraining distribution.

[8] von Oswald, Johannes, et al. "Transformers Learn In-Context by Gradient Descent." *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[9] Wu, Shiguang, et al. "Why In-Context Learning Models are Good Few-Shot Learners?" *International Conference on Learning Representations*, 2025.

ICL as Meta Optimizer

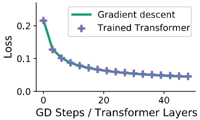


Figure: SGD-Transformer Equivalence

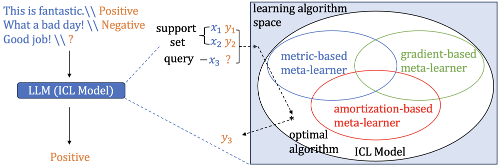


Figure: ICL as a Meta-Optimizer

Meta-Learning View: ICL acts as data-dependent meta-learning, distinct from gradient-/metric-/amortized meta-learners. ⇒ Implicit algorithm is shaped by pretraining distribution.[8] [9]

[8] von Oswald, Johannes, et al. "Transformers Learn In-Context by Gradient Descent." *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[9] Wu, Shiguang, et al. "Why In-Context Learning Models are Good Few-Shot Learners?" *International Conference on Learning Representations*, 2025.

ICL as SGD?

- *Lingfeng et al* questioned the ICL $\approx$ SGD approach.

[10] Shen, Lingfeng. "Do Pretrained Transformers Learn In-Context by Gradient Descent?" *Proceedings of the 41st International Conference on Machine Learning*, 2024.

ICL as SGD?

- *Lingfeng et al* questioned the ICL $\approx$ SGD approach.
- They questioned both the **Emergent equivalence** and the **Constructive Equivalence**

[10] Shen, Lingfeng. "Do Pretrained Transformers Learn In-Context by Gradient Descent?" *Proceedings of the 41st International Conference on Machine Learning*, 2024.

ICL as SGD?

- *Lingfeng et al* questioned the ICL $\approx$ SGD approach.
- They questioned both the **Emergent equivalence** and the **Constructive Equivalence**
-

$$\underbrace{M_{\Theta_0}(\sigma_A \circ x_t) - M_{\Theta_0}(\sigma_B \circ x_t)}_{\text{The order sensitivity of ICL}} = \underbrace{M_{\Theta_{\sigma_A}}(x_t) - M_{\Theta_{\sigma_B}}(x_t)}_{\text{The order sensitivity of algorithm } A}$$

[10] Shen, Lingfeng. "Do Pretrained Transformers Learn In-Context by Gradient Descent?" *Proceedings of the 41st International Conference on Machine Learning*, 2024.

ICL as SGD?

- *Lingfeng et al* questioned the ICL $\approx$ SGD approach.
- They questioned both the **Emergent equivalence** and the **Constructive Equivalence**
-

$$\underbrace{M_{\Theta_0}(\sigma_A \circ x_t) - M_{\Theta_0}(\sigma_B \circ x_t)}_{\text{The order sensitivity of ICL}} = \underbrace{M_{\Theta_{\sigma_A}}(x_t) - M_{\Theta_{\sigma_B}}(x_t)}_{\text{The order sensitivity of algorithm } A}$$

- Across all metrics(Accuracy, Cosine similarity among others), ICL and GD show inconsistent behavior.

[10] Shen, Lingfeng. "Do Pretrained Transformers Learn In-Context by Gradient Descent?" *Proceedings of the 41st International Conference on Machine Learning*, 2024.

ICL as SGD?

- *Lingfeng et al* questioned the ICL $\approx$ SGD approach.
- They questioned both the **Emergent equivalence** and the **Constructive Equivalence**
-

$$\underbrace{M_{\Theta_0}(\sigma_A \circ x_t) - M_{\Theta_0}(\sigma_B \circ x_t)}_{\text{The order sensitivity of ICL}} = \underbrace{M_{\Theta_{\sigma_A}}(x_t) - M_{\Theta_{\sigma_B}}(x_t)}_{\text{The order sensitivity of algorithm } A}$$

- Across all metrics(Accuracy, Cosine similarity among others), ICL and GD show inconsistent behavior. [10]

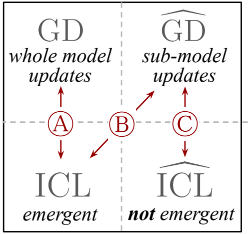


Figure: ICL GD Equivelence

[10] Shen, Lingfeng. "Do Pretrained Transformers Learn In-Context by Gradient Descent?" *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Hypothesis 2: ICL as Bayesian Inference

ICL and Bayesian Inference

$$\boxed{p(\text{output} \mid \text{prompt}) = \int_{\text{concept}} p(\text{output} \mid \text{concept}, \text{prompt})\, p(\text{concept} \mid \text{prompt})\, d(\text{concept})}$$

[11] Ahuja, Kabir, et al. "In-Context Learning through the Bayesian Prism." *The Twelfth International Conference on Learning Representations*, 2024.

ICL and Bayesian Inference

$$\boxed{p(\text{output} \mid \text{prompt}) = \int_{\text{concept}} p(\text{output} \mid \text{concept}, \text{prompt})\, p(\text{concept} \mid \text{prompt})\, d(\text{concept})}$$

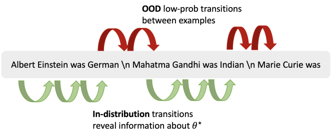


Figure: Signal and the OOD [a]

[a] Xie, Sang Michael, et al. "An Explanation of In-context Learning as Implicit Bayesian Inference." *The Tenth International Conference on Learning Representations, ICLR 2022*, 2022.

[11] Ahuja, Kabir, et al. "In-Context Learning through the Bayesian Prism." *The Twelfth International Conference on Learning Representations*, 2024.

ICL and Bayesian Inference

$$p(\text{output} \mid \text{prompt}) = \int_{\text{concept}} p(\text{output} \mid \text{concept}, \text{prompt})\, p(\text{concept} \mid \text{prompt})\, d(\text{concept})$$

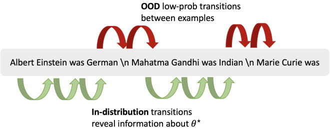


Figure: Signal and the OOD [a]

Hierarchical Meta ICL Setup

$$\begin{aligned}
&c \sim \pi = (\pi_1, \ldots, \pi_M),\\
&f \sim p(f \mid c),\\
&(x_i, y_i) :\ x_i \sim p(x),\ y_i = f(x_i),\ i = 1, \ldots, N,\\
&x_{N+1}\text{: query input},\ y_{N+1}\text{: to predict.}
\end{aligned}$$

Bayes optimal inference:

$$\Pr(c \mid \{(x_i, y_i)\}_{i=1}^{N}) \propto \pi_c \prod_{i=1}^{N} \Pr(y_i \mid x_i, c),$$

$$\Pr(y_{N+1} \mid x_{N+1}, \text{context}) = \sum_{c=1}^{M} \Pr(c \mid \text{context})\, \mathbb{E}_{f\sim p(\cdot \mid c)} \left[\Pr(y_{N+1} \mid x_{N+1}, f)\right].$$

[a] Xie, Sang Michael, et al. "An Explanation of In-context Learning as Implicit Bayesian Inference." *The Tenth International Conference on Learning Representations, ICLR 2022*, 2022.

11

[11] Ahuja, Kabir, et al. "In-Context Learning through the Bayesian Prism." *The Twelfth International Conference on Learning Representations*, 2024.

ICL and Bayesian Inference

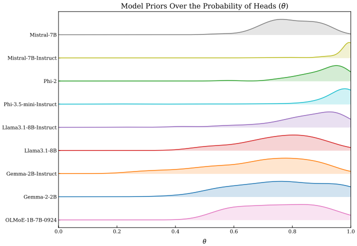


Figure: Coin Prior

[12] Gupta, Ritwik, et al. "Enough Coin Flips Can Make LLMs Act Bayesian." *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2025.

ICL and Bayesian Inference

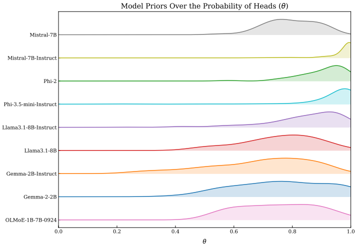


Figure: Coin Prior

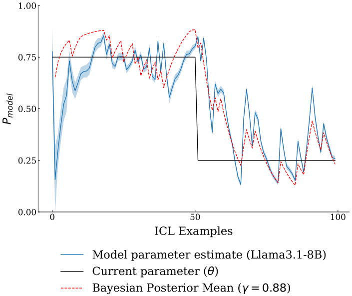


Figure: Posterior

12

[12] Gupta, Ritwik, et al. "Enough Coin Flips Can Make LLMs Act Bayesian." *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2025.

ICL and Bayesian Inference

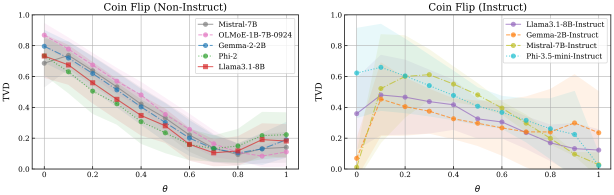


Figure: Biased Coin Instruct vs Non-Instruct

[13]

1. LLMs have biased *priors*.
2. Initial predictions diverge from ground truth due to these.
3. Explicit biasing (using prompts) improves only Instruct LLMs.
4. ICL helps remove the bias, similar to *Bayesian Updates*.

[13] Gupta, Ritwik, et al. ”Enough Coin Flips Can Make LLMs Act Bayesian.” *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2025.

Hypothesis 3: ICL and Induction Heads

Induction Heads

-

$$[A^*][B^*]\dots[A] \to [B]$$

where $A^* \approx A$ and $B^* \approx B$ are similar in some space.

[14] Olsson, Catherine, et al. "In-context Learning and Induction Heads." *CoRR, abs/2209.11895*, 2022.

Induction Heads

- $$[A^*][B^*]\dots[A] \to [B]$$

  where $A^* \approx A$ and $B^* \approx B$ are similar in some space.

- They **emerge** during training alongside ICL ability.

[14] Olsson, Catherine, et al. "In-context Learning and Induction Heads." *CoRR, abs/2209.11895*, 2022.

Induction Heads

- $$[A^*][B^*]\dots[A] \to [B]$$

  where $A^* \approx A$ and $B^* \approx B$ are similar in some space.

- They **emerge** during training alongside ICL ability.
- Core mechanism behind ICL.

[14] Olsson, Catherine, et al. "In-context Learning and Induction Heads." *CoRR, abs/2209.11895*, 2022.

Induction Heads

- $$[A^*][B^*]\dots[A] \to [B]$$

  where $A^* \approx A$ and $B^* \approx B$ are similar in some space.

- They **emerge** during training alongside ICL ability.
- Core mechanism behind ICL.
- Ablation shows they are **causal** for ICL in small models.

[14] Olsson, Catherine, et al. "In-context Learning and Induction Heads." *CoRR, abs/2209.11895*, 2022.

Induction Heads

- $$[A^*][B^*]\dots[A] \rightarrow [B]$$

  where $A^* \approx A$ and $B^* \approx B$ are similar in some space.

- They **emerge** during training alongside ICL ability.
- Core mechanism behind ICL.
- Ablation shows they are **causal** for ICL in small models.
- Found across various model sizes and tasks.

[14] Olsson, Catherine, et al. "In-context Learning and Induction Heads." *CoRR, abs/2209.11895*, 2022.

Induction Heads

- $$[A^*][B^*]\dots[A] \rightarrow [B]$$

  where $A^* \approx A$ and $B^* \approx B$ are similar in some space.

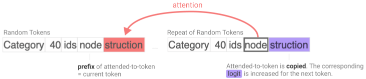


Figure: Induction Heads

- They **emerge** during training alongside ICL ability.
- Core mechanism behind ICL.
- Ablation shows they are **causal** for ICL in small models.
- Found across various model sizes and tasks.

[14] Olsson, Catherine, et al. "In-context Learning and Induction Heads." *CoRR, abs/2209.11895*, 2022.

Induction Heads

- $$[A^*][B^*]\dots[A] \rightarrow [B]$$

  where $A^* \approx A$ and $B^* \approx B$ are similar in some space.

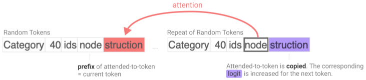


Figure: Induction Heads

- They **emerge** during training alongside ICL ability.
- Core mechanism behind ICL.
- Ablation shows they are **causal** for ICL in small models.
- Found across various model sizes and tasks.

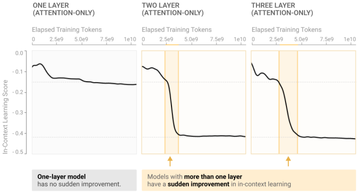


Figure: Abrupt Loss Transition

14

[14] Olsson, Catherine, et al. "In-context Learning and Induction Heads." *CoRR, abs/2209.11895*, 2022.

Learning Plateau's and Abrupt Switching of ICL

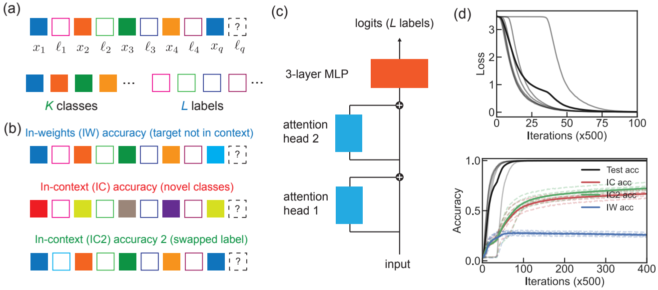


Figure: Interpreting In-Context Classification Task

15

[15] Reddy, Gautam. "The Mechanistic Basis of Data Dependence and Abrupt Learning in an In-Context Classification Task." *International Conference on Learning Representations*, 2024.

Learning Plateau's and Abrupt Switching of ICL

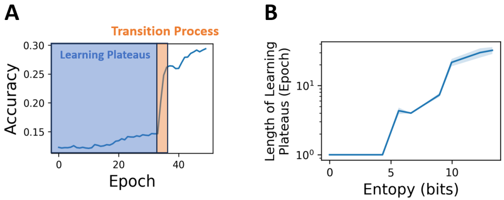


Figure: Plateau in ICL

$\Rightarrow$ **Burstiness**, **Large Vocabulary Size**, **Skewed Classes** and **High Diversity** within Class promote ICL. [a]

$\Rightarrow$ Decompose Representation from parameters (W) and parameters + context (C).

$\Rightarrow$ Transferring Embeddings and Initial layers eliminates plateaus. [b]

---

[a] Reddy, Gautam. "The Mechanistic Basis of Data Dependence and Abrupt Learning in an In-Context Classification Task." *International Conference on Learning Representations*, 2024.

[b] Fu, Jingwen, et al. "Breaking through the Learning Plateaus of In-context Learning in Transformer." *Proceedings of the 41st International Conference on Machine Learning*, 2024.

ICL Across Architectures

- *Lee et al* conduct an empirical study comparing ICL performance across diverse model architectures such as **CNNs**, **RNNs**, **Transformers** and **SSMs**.

[16] Lee, Ivan, et al. ”Is Attention Required for ICL? Exploring the Relationship Between Model Architecture and In-Context Learning Ability.” *The Twelfth International Conference on Learning Representations*, 2024.

[17] Park, Jongho, et al. ”Can Mamba Learn How to Learn? A Comparative Study on In-Context Learning Tasks.” *arXiv preprint arXiv:2402.04248*, 2024.

ICL Across Architectures

- *Lee et al* conduct an empirical study comparing ICL performance across diverse model architectures such as **CNNs**, **RNNs**, **Transformers** and **SSMs**.
- All considered architectures achieve ICL, and some attention alternatives not only match but even surpass transformers.

[16] Lee, Ivan, et al. "Is Attention Required for ICL? Exploring the Relationship Between Model Architecture and In-Context Learning Ability." *The Twelfth International Conference on Learning Representations*, 2024.

[17] Park, Jongho, et al. "Can Mamba Learn How to Learn? A Comparative Study on In-Context Learning Tasks." *arXiv preprint arXiv:2402.04248*, 2024.

ICL Across Architectures

- *Lee et al* conduct an empirical study comparing ICL performance across diverse model architectures such as **CNNs**, **RNNs**, **Transformers** and **SSMs**.
- All considered architectures achieve ICL, and some attention alternatives not only match but even surpass transformers.

[16] [17]

Task	Prompt	Target	
Associative Recall	a, 1, b, 3, c, 2, b	3	
Linear Regression	$\mathbf{x}_1, y_1, \mathbf{x}_2, y_2, \mathbf{x}_3, y_3, \mathbf{x}_4$	y_4	$\exists \mathbf{w}$ such that $\forall i, y_i = \mathbf{x}_i \cdot \mathbf{w}$
Multiclass Classification	$\mathbf{x}_1$, b, $\mathbf{x}_2$, a, $\mathbf{x}_3$, a, $\mathbf{x}_4$	b	$x_1, x_4 \sim \mathcal{N}(y_b, I_d)$ $x_2, x_3 \sim \mathcal{N}(y_a, I_d)$
Image Classification		4	bursty training prompt
		2	non-bursty training prompt
		0	evaluation prompt
Language Modeling	*Colorless green ideas sleep*	*furiously*	

[16] Lee, Ivan, et al. "Is Attention Required for ICL? Exploring the Relationship Between Model Architecture and In-Context Learning Ability." *The Twelfth International Conference on Learning Representations*, 2024.

[17] Park, Jongho, et al. "Can Mamba Learn How to Learn? A Comparative Study on In-Context Learning Tasks." *arXiv preprint arXiv:2402.04248*, 2024.

ICL Across Architectures

- *Tong et al* test the ICL capabilities in **MLPs** and **MLP-Mixer** architectures.

[18]Tong, William L., and Cengiz Pehlevan. ”MLPs Learn In-Context on Regression and Classification Tasks.” *The Thirteenth International Conference on Learning Representations*, 2025.

ICL Across Architectures

- *Tong et al* test the ICL capabilities in **MLPs** and **MLP-Mixer** architectures.
- Experimentally, they observe that given sufficient compute, MLP based achitectures perform comparably with the transformers.

[18] Tong, William L., and Cengiz Pehlevan. ”MLPs Learn In-Context on Regression and Classification Tasks.” *The Thirteenth International Conference on Learning Representations*, 2025.

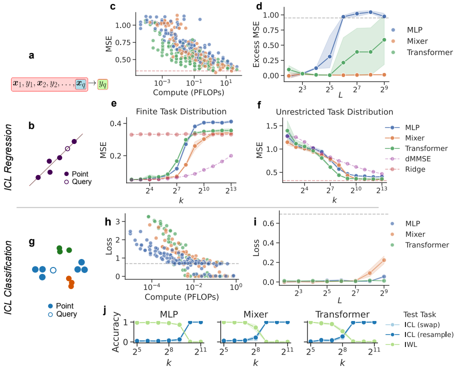
ICL Across Architectures

- *Tong et al* test the ICL capabilities in **MLPs** and **MLP-Mixer** architectures.
- Experimentally, they observe that given sufficient compute, MLP based achitectures perform comparably with the transformers.
- Also, for tasks which align with the architecture bias, MLP based architectures even outperform the transformers.

[18] Tong, William L., and Cengiz Pehlevan. "MLPs Learn In-Context on Regression and Classification Tasks." *The Thirteenth International Conference on Learning Representations*, 2025.

ICL Across Architectures

- *Tong et al* test the ICL capabilities in **MLPs** and **MLP-Mixer** architectures.
- Experimentally, they observe that given sufficient compute, MLP based achitectures perform comparably with the transformers.
- Also, for tasks which align with the architecture bias, MLP based architectures even outperform the transformers.



18

---

18 Tong, William L., and Cengiz Pehlevan. "MLPs Learn In-Context on Regression and Classification Tasks." *The Thirteenth International Conference on Learning Representations*, 2025.

Thank you!

This presentation is typeset using the LaTeX beamer package with the
UIC Presentation Template:
`https://www.overleaf.com/latex/templates/uic-presentation-template/dgjbtyvtgqcg`