

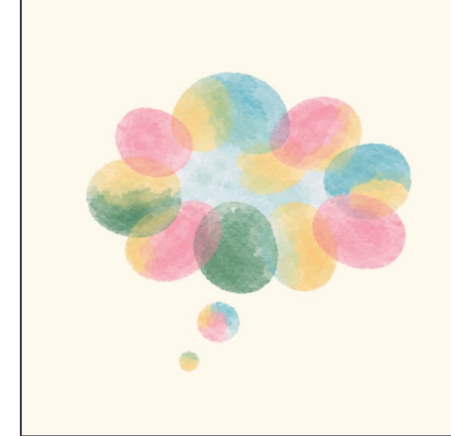
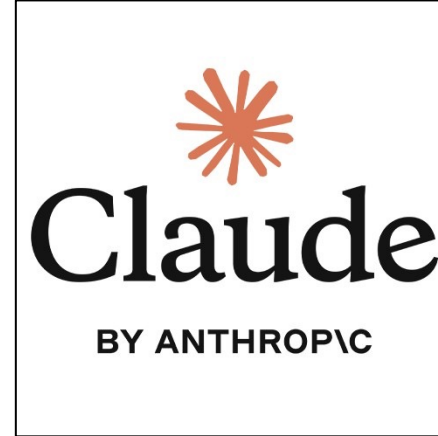
Growing Reasoning Models: Some Progress and More Questions

Daniel Khashabi



The success we dreamed of

Language models that **reason**
across many tasks that were once considered aspirational.



The success we dreamed of

Let n be an even positive integer. Let p be a monic, real polynomial of degree $2n$; that is to say,

$$p(x) = x^{2n} + a_{2n-1}x^{2n-1} + \cdots + a_1x + a_0$$

for some real coefficients a_0, \dots, a_{2n-1} . Suppose that $p\left(\frac{1}{k}\right) = k^2$ for all integers k such that $1 \leq |k| \leq n$.

Find all other real numbers x for which $p\left(\frac{1}{x}\right) = x^2$.

The success we dreamed of

Let n be an even positive integer. Let p be a monic, real polynomial of degree $2n$; that is to say,

$$p(x) = x^{2n} + a_{2n-1}x^{2n-1} + \cdots + a_1x + a_0$$

for some real coefficients a_0, \dots, a_{2n-1} . Suppose that $p\left(\frac{1}{k}\right) = k^2$ for all integers k such that $1 \leq |k| \leq n$.

Find all other real numbers x for which $p\left(\frac{1}{x}\right) = x^2$.



Thought for 5 seconds ▾

The success we dreamed of

Let n be an even positive integer. Let p be a monic, real polynomial of degree $2n$; that is to say,

$$p(x) = x^{2n} + a_{2n-1}x^{2n-1} + \cdots + a_1x + a_0$$

for some real coefficients a_0, \dots, a_{2n-1} . Suppose that $p\left(\frac{1}{k}\right) = k^2$ for all integers k such that $1 \leq |k| \leq n$.

Find all other real numbers x for which $p\left(\frac{1}{x}\right) = x^2$.



Thought for 5 seconds ▾

To solve this problem, we need to find all real numbers x such that

$$p\left(\frac{1}{x}\right) = x^2,$$

where p is a monic real polynomial of degree $2n$, and it is given that

$$p\left(\frac{1}{k}\right) = k^2$$

for all integers k with $1 \leq |k| \leq n$.

Understanding the Given Information:

- 1. Given Data Points:** We have $2n$ data points from $k = \pm 1, \pm 2, \dots, \pm n$, which provide the values $p\left(\frac{1}{k}\right) = k^2$.
- 2. Polynomial Function:** $p(x)$ is monic of degree $2n$, which means its highest-degree coefficient is 1.

- The leading term of $s(x)$ is $-x^{2n+2}$.
- The leading term of $m(x)h(x)$ is $(x^{2n})(-x^2) = -x^{2n+2}$, ensuring the degrees match.
- Comparing the constant term of $s(x)$ using $x = 0$:
- $s(0) = p(0) - 0 = 1$, where $p(x) = x^{2n} p\left(\frac{1}{x}\right)$.
- Evaluating $m(0)$ and $h(0)$:
- $m(0) = (-1)^n (n!)^2$.
- $h(0) = -0 + c = c$.
- Since $s(0) = m(0)h(0)$:

$$1 = (-1)^n (n!)^2 c,$$

which gives:

$$c = \frac{1}{(-1)^n (n!)^2} = \frac{(-1)^n}{(n!)^2}.$$

- For n even (since n is even), $(-1)^n = 1$, so $c > 0$.

Finding Additional Solutions:

Solving $h(x) = 0$:

$$-x^2 + c = 0 \implies x^2 = c \implies x = \pm \frac{1}{n!},$$

since $c = \frac{1}{(n!)^2}$.

Conclusion:

The additional real numbers x satisfying $p\left(\frac{1}{x}\right) = x^2$ are $x = \pm \frac{1}{n!}$.

- The leading term of $s(x)$ is $-x$.
- The leading term of $m(x)h(x)$ is $(x^{2n})(-x^2) = -x^{2n+2}$, ensuring the degrees match.
- Comparing the constant term of $s(x)$ using $x = 0$:
- $s(0) = p(0) - 0 = 1$, where $p(x) = x^{2n}$.
- Evaluating $m(0)$ and $h(0)$:
- $m(0) = (-1)^n (n!)^2$.
- $h(0) = -0 + c = c$.
- Since $s(0) =$

$$1 = (-1)^n (n!)^2 c$$

which gives:

$$c = \frac{1}{(-1)^n (n!)^2} =$$

- For n even (since n is even), $(-1)^n = 1$, so $c > 0$.

How did we get here?

Finding Additional Solutions:

Solving $h(x) = 0$:

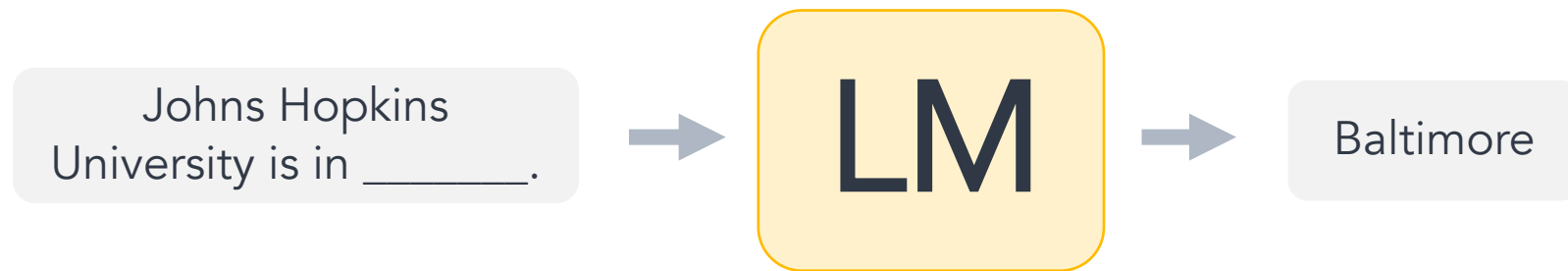
$$-x^2 + c = 0 \implies x^2 = c \implies x = \pm \frac{1}{n!},$$

since $c = \frac{1}{(n!)^2}$.

Conclusion:

The additional real numbers x satisfying $p\left(\frac{1}{x}\right) = x^2$ are $x = \pm \frac{1}{n!}$.

Reasoning Stack: (1) Next-Token Prediction



Reasoning Stack: (2) Post-training

Reasoning Stack: (2) Post-training

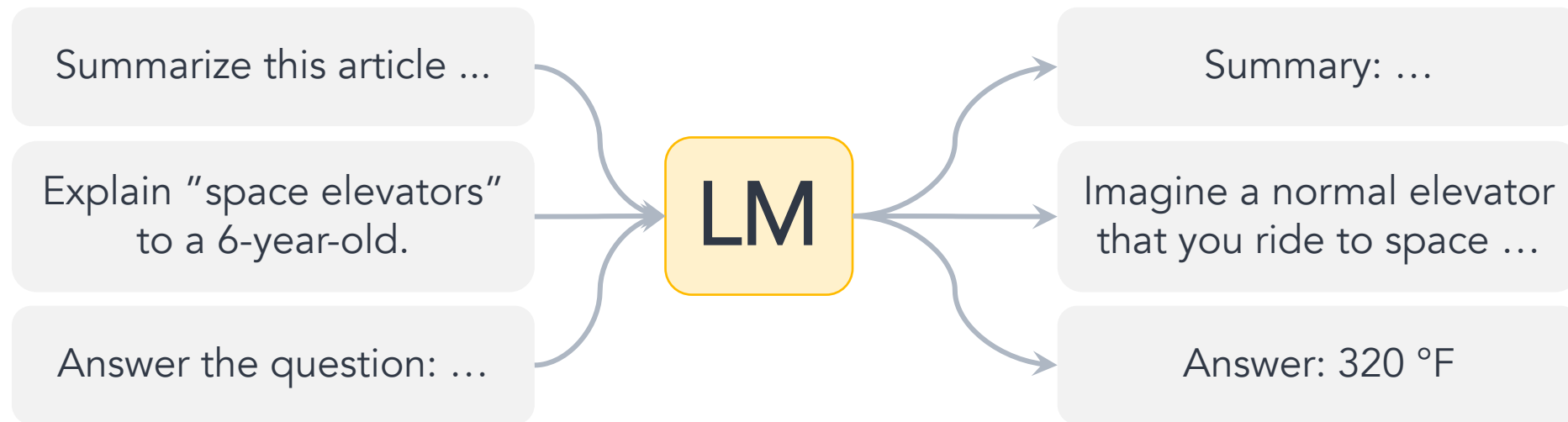
- “Aligning” LMs to respect our intents and instructions.
 - Supervised Fine-tuning (SFT/behavior cloning) with labeled data.

Cross-Task Generalization via Natural Language Crowdsourcing Instructions.
Mishra, Khashabi, Baral and Hajishirzi. *ACL* 2021 🏆 **Ai2 Lasting Impact Award!** 🏆

Among others: Sanh et al. 2022; Chung et al. 2022,

Reasoning Stack: (2) Post-training

- “Aligning” LMs to respect our intents and instructions.
 - Supervised Fine-tuning (SFT/behavior cloning) with labeled data.



Cross-Task Generalization via Natural Language Crowdsourcing Instructions.
Mishra, Khashabi, Baral and Hajishirzi. *ACL 2021* 🏆 **Ai2 Lasting Impact Award!** 🏆

Among others: Sanh et al. 2022; Chung et al. 2022,

Reasoning Stack: (2) Post-training

- “Aligning” LMs to respect our intents and instructions.
 - Supervised Fine-tuning (SFT/behavior cloning) with labeled data.

Cross-Task Generalization via Natural Language Crowdsourcing Instructions.
Mishra, Khashabi, Baral and Hajishirzi. *ACL* 2021 🏆 **Ai2 Lasting Impact Award!** 🏆

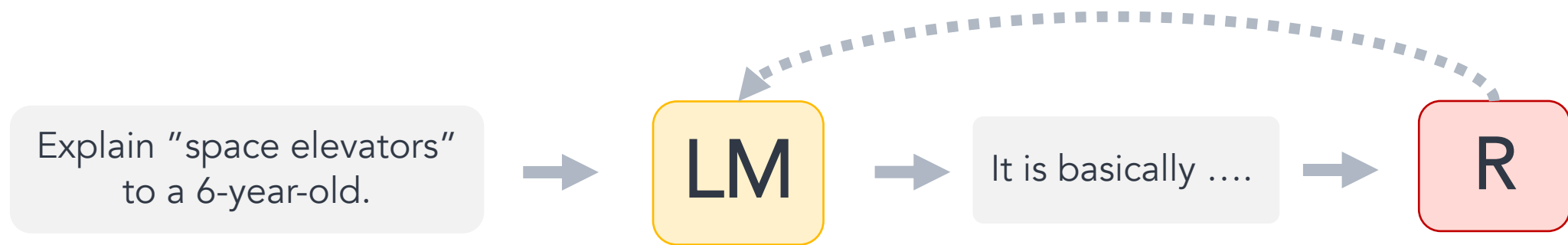
Among others: Sanh et al. 2022; Chung et al. 2022,

Reasoning Stack: (2) Post-training

- “Aligning” LMs to respect our intents embedded in instructions.
 - Supervised Fine-tuning (SFT/behavior cloning) with labeled data.
 - Reinforcement Learning on preference data or verifiers.

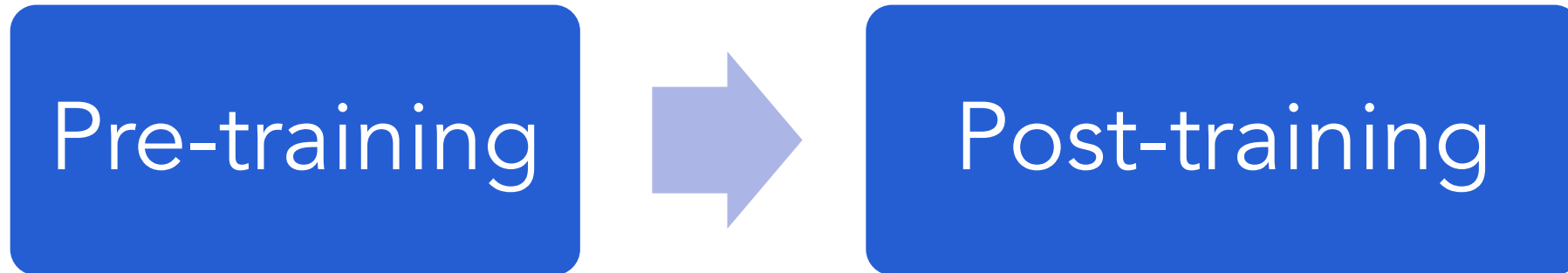
Reasoning Stack: (2) Post-training

- “Aligning” LMs to respect our intents embedded in instructions.
 - Supervised Fine-tuning (SFT/behavior cloning) with labeled data.
 - Reinforcement Learning on preference data or verifiers.



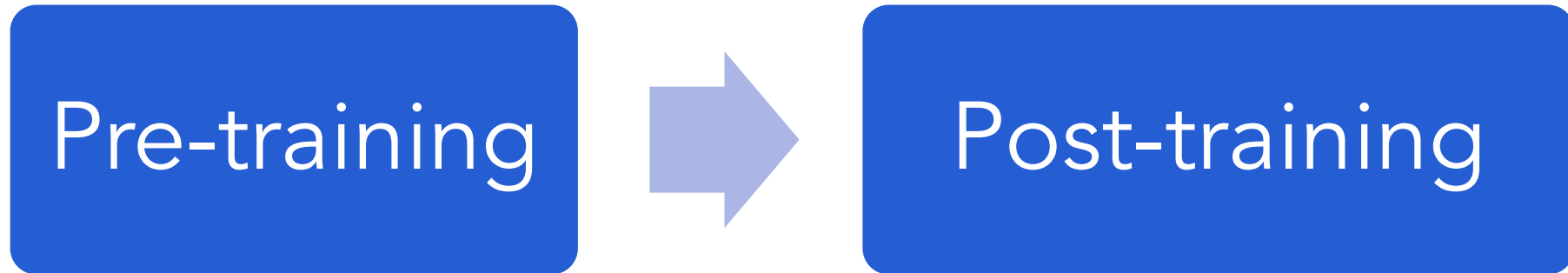
Reasoning Stack

- Almost all the reasoning models follow this recipe:



Rethinking the Reasoning Stack

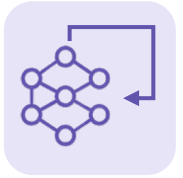
- Almost all the reasoning models follow this recipe:
- Reasoning is **not** limited to post-training. It's the whole stack.



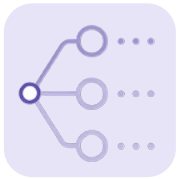
Today: Rethinking the Reasoning Stack



1. How does in-context learning (ICL) **emerge**? (pre-training)



2. How do we **distill ICL**? (pre \rightarrow post)



3. How do we maintain **diversity**? (post-training)

Today: Rethinking the Reasoning Stack



1. How does in-context learning (ICL) **emerge**? (pre-training)



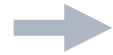
2. How do we **distill ICL**? (pre → post)



3. How do we maintain **diversity**? (post-training)

In-Context Learning

Input: NYU Output: NYC
Input: UMD Output: DC
Input: JHU Output:



Baltimore

In-Context Learning

- **ICL** := learning to imitate the *implicit* pattern described by few examples provided in the context.

Input: NYU Output: NYC
Input: UMD Output: DC
Input: JHU Output:



LM



Baltimore

- Nearly all modern LLMs use [few-shot] ICL prompts to track downstream performance during pre-training.
GPT3, LLaMA family, Mistral, Gemma, DeepSeek, etc.

In-Context Learning

- **ICL** := learning to imitate the *implicit* pattern described by few examples provided in the context.

Input: NYU Output: NYC
Input: UMD Output: DC
Input: JHU Output:



LM

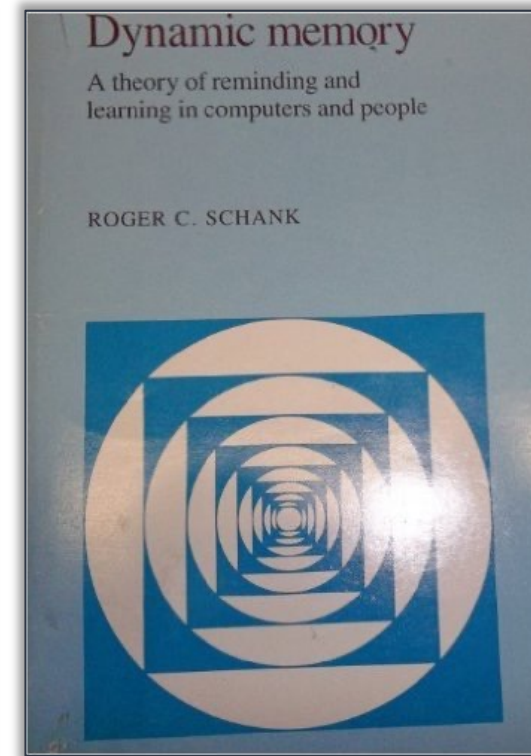
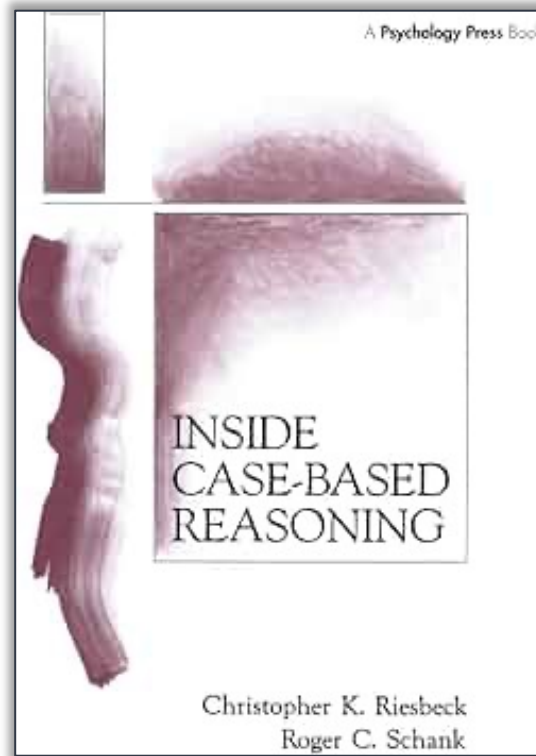
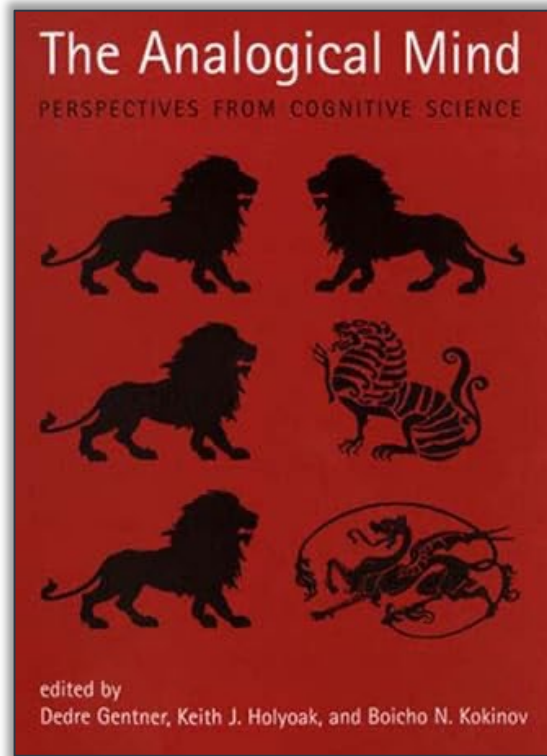


Baltimore

- Nearly all modern LLMs use [few-shot] ICL prompts to track downstream performance during pre-training.
GPT3, LLaMA family, Mistral, Gemma, DeepSeek, etc.

ICL is intellectually intriguing

An organic form of *reasoning* mechanism we've been looking for years!



Analogical reasoning, case-based reasoning, inductive learning, ...

The big **open** question:
Why (and When) does ICL emerge?

The big **open** question:
Why (and When) does ICL emerge?

- For years, the organic emergence ICL (from 'next-token' training) is deeply tied to **human language**.
- Perhaps human language has specific properties that yields ICL.
 - Compositionality, abstraction, parallel patterns, etc.

A Theory of Emergent In-Context Learning as Implicit
Structure Induction

Michael Hahn
Saarland University

hahn@iit.uni-saarland.de

Navin Goyal
Microsoft Research India
navingo@microsoft.com

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

Yanda Chen¹ Chen Zhao^{2,3} Zhou Yu¹ Kathleen McKeown¹ He He²

¹Columbia University, ²New York University, ³NYU Shanghai

{yanda.chen, kathy}@cs.columbia.edu, cz1285@nyu.edu
zy2461@columbia.edu, hehe@cs.nyu.edu

The big **open** question:
Why (and When) does ICL emerge?

- For years, the organic emergence ICL (from 'next-token' training) is deeply tied to **human language**.
- Perhaps human language has specific properties that yields ICL.
 - Compositionality, abstraction, parallel patterns, etc.

A Theory of Emergent In-Context Learning as Implicit
Structure Induction

Michael Hahn
Saarland University

hahn@iit.uni-saarland.de

Navin Goyal
Microsoft Research India
navingo@microsoft.com

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

Yanda Chen¹ Chen Zhao^{2,3} Zhou Yu¹ Kathleen McKeown¹ He He²

¹Columbia University, ²New York University, ³NYU Shanghai

{yanda.chen, kathy}@cs.columbia.edu, cz1285@nyu.edu
zy2461@columbia.edu, hehe@cs.nyu.edu

The big **open** question:

Why (and When) does ICL emerge?

- For years, the organic emergence ICL (from 'next-token' training) is deeply tied to **human language**.
- But ... is it really about human language specifically?

The big **open** question:

Why (and When) does ICL emerge?

- For years, the organic emergence ICL (from 'next-token' training) is deeply tied to **human language**.
- But ... is it really about human language specifically?

Research question: Is ICL a [human] language-specific phenomenon?

Human vs. Genetic Language

Genomic Next-Token Predictors are In-Context Learners.
Beslow, Mishra, Revsine, Schatz, Liu, Khashabi. TMLR. 2026.



Human vs. Genetic Language

Human language

The computer science department at JHU is known for ...



Human vs. Genetic Language

Human language

The computer science department at JHU is known for ...

Genomic language

ATTACCCCGATTGCTATGCCTGAGAAGCTATTATGCCTGAGAAGCTAT ...



Human vs. Genetic Language

Human language

The computer science department at JHU is known for ...

Genomic language

ATTACCCCGATTGCTATGCCTGAGAAGCTATTATGCCTGAGAAGCTAT ...

- **Similar:** Both are symbolic, short- and long-range dependencies.
- **Different:** Vocabularies, interpretability, evolutionary path.



Human vs. Genetic Language

Human language

The computer science department at JHU is known for ...

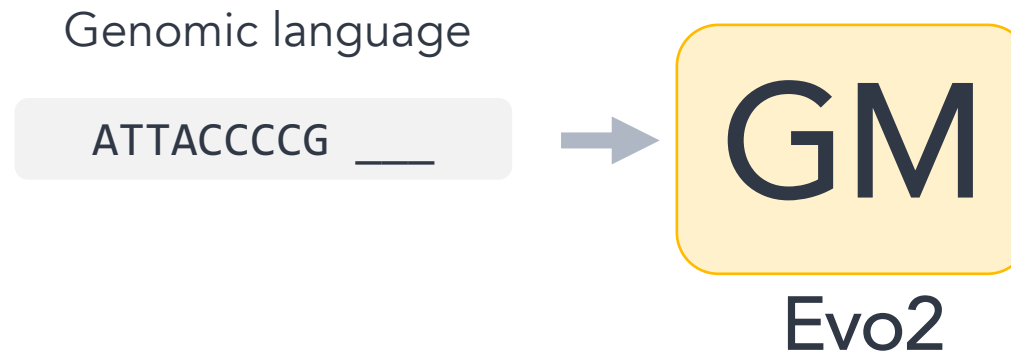
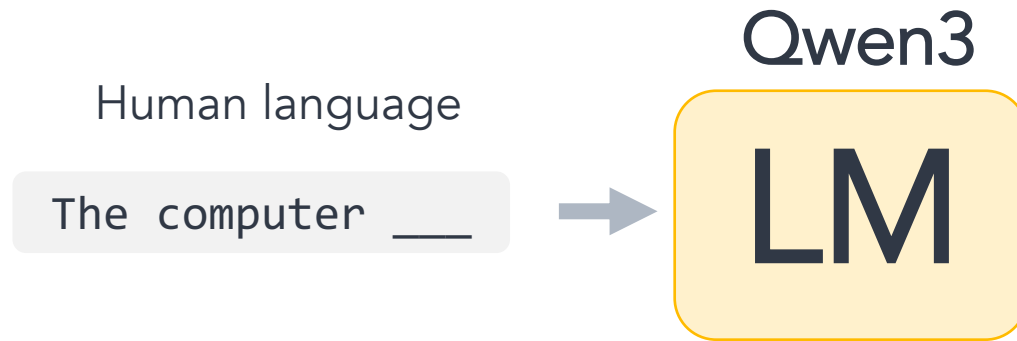
Genomic language

ATTACCCCGATTGCTATGCCTGAGAAGCTATTATGCCTGAGAAGCTAT ...

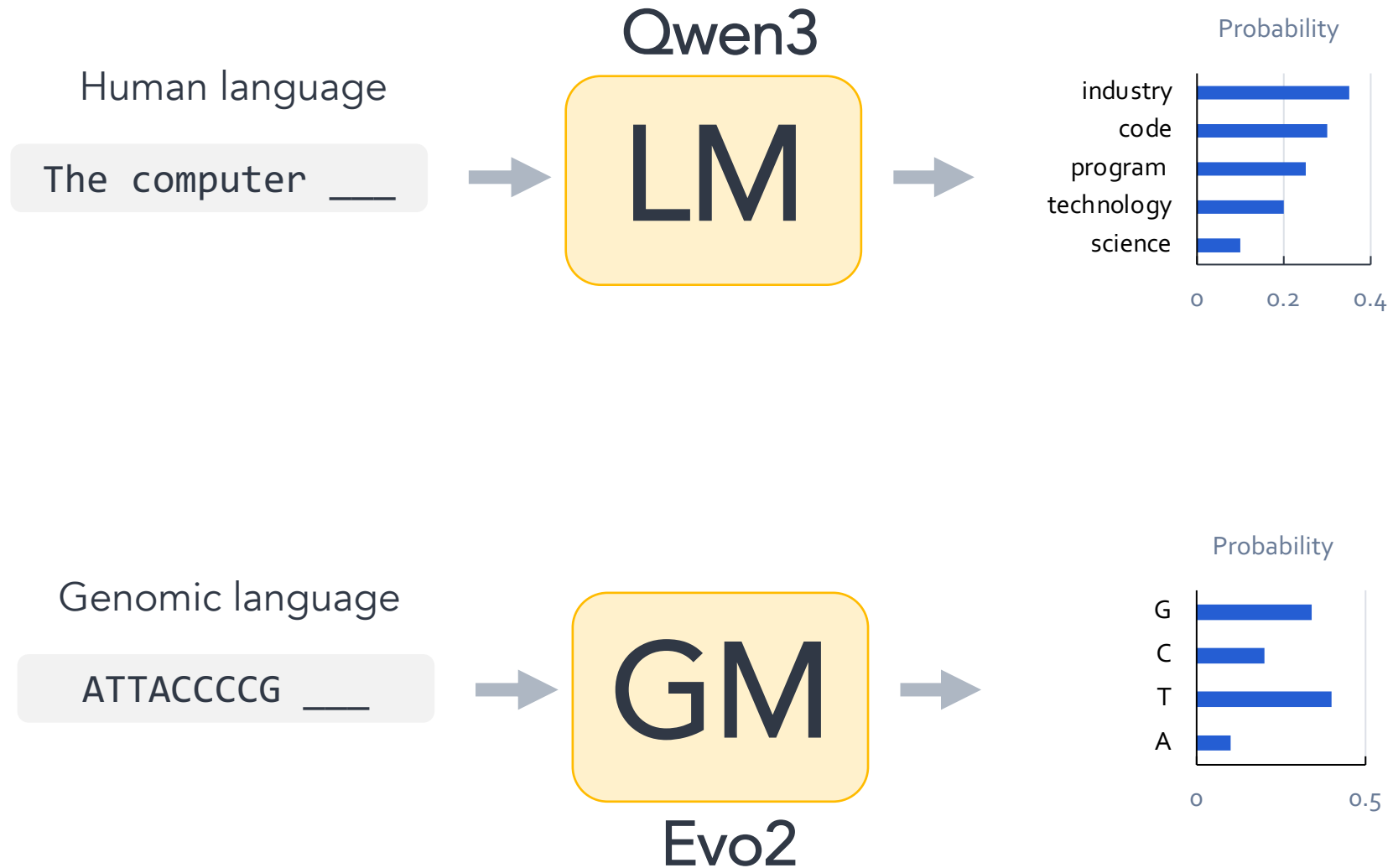
- **Similar:** Both are symbolic, short- and long-range dependencies.
- **Different:** Vocabularies, interpretability, evolutionary path.



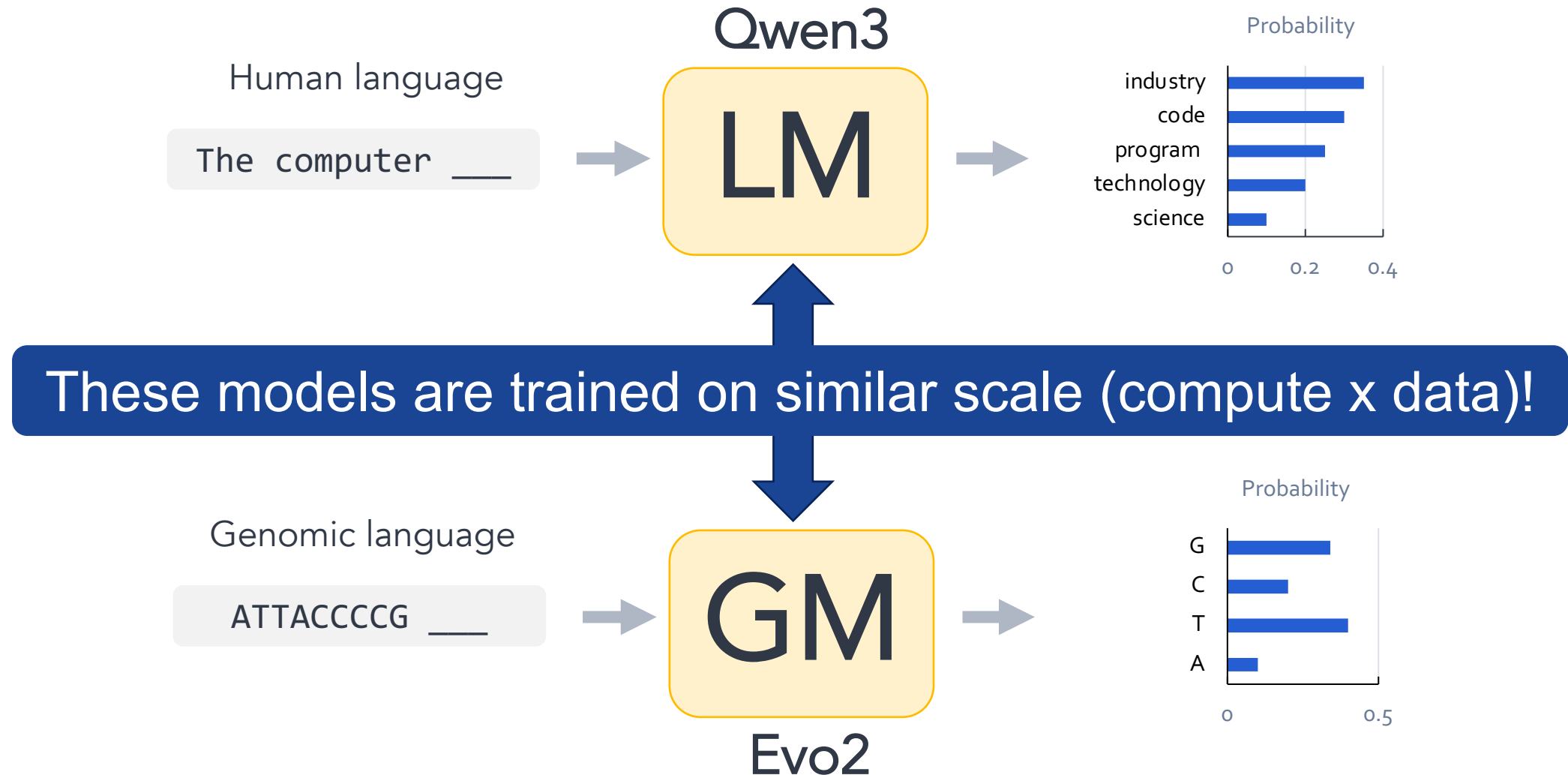
ICL on Human vs. Genetic Language



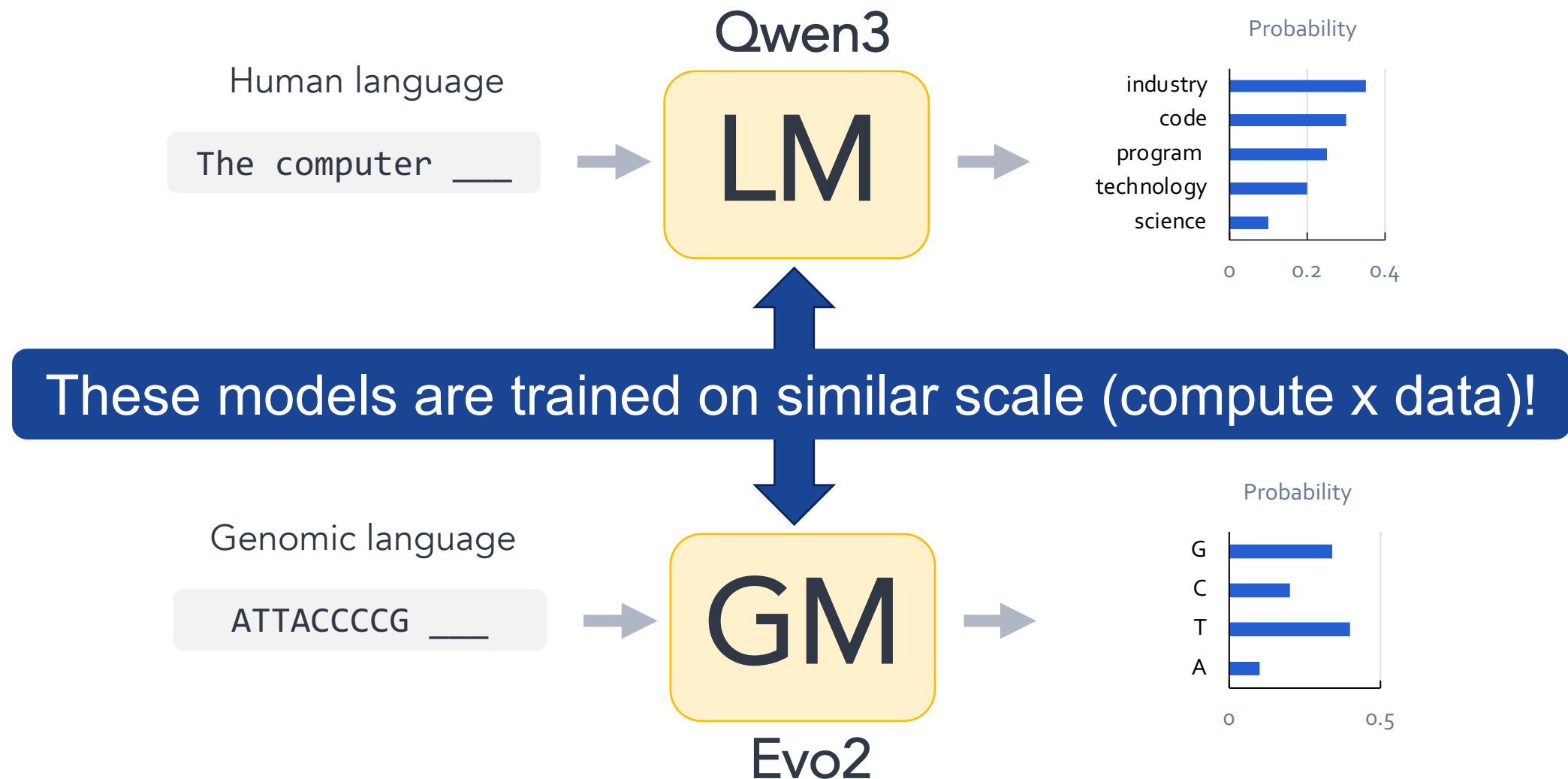
ICL on Human vs. Genetic Language



ICL on Human vs. Genetic Language



Hypothesis: If ICL appears in both settings, it must reflect a *more general computational mechanism*.



What task should we use for evaluation?

```
input1 -> Output1 SEP
input2 -> Output2 SEP
input3 -> Output3 SEP
Input4 -> ?
```



What task should we use for evaluation?

We defined 100 reasoning tasks based on bitstrings

```
10100000 -> 00000101 SEP
11100011 -> 11000111 SEP
11001110 -> 01110011 SEP
11000000 -> ?
```

Various functions: Bitwise NOT, Reverse, etc.



What task should we use for evaluation?

We defined 100 reasoning tasks based on bitstrings

```
10100000 -> 00000101 SEP
11100011 -> 11000111 SEP
11001110 -> 01110011 SEP
11000000 -> ?
```

Various functions: Bitwise NOT, Reverse, etc.

But we need to express the *same underlying task* in a language each model understands.



What task should we use for evaluation?

```
10100000 -> 00000101 SEP  
11100011 -> 11000111 SEP  
11001110 -> 01110011 SEP  
11000000 -> ?
```



What task should we use for evaluation?

```
101000000 -> 00000101 SEP
111000011 -> 11000111 SEP
11001110 -> 01110011 SEP
110000000 -> ?
```

- Replace "1" with "T"
- Replace "0" with "A"
- Replace "SEP" with "G"
- Drop "->"



What task should we use for evaluation?



Genomic Next-Token Predictors are In-Context Learners. **Evo2**
Beslow, Mishra, Revsine, Schatz, Liu, Khashabi. TMLR. 2026.

What task should we use for evaluation?

- Replace "1" with "3"
- Replace "0" with "4"
- Replace "SEP" with "6"
- Drop "->"

```
10100000 -> 00000101 SEP
11100011 -> 11000111 SEP
11001110 -> 01110011 SEP
11000000 -> ?
```

```
34344444444444443436333
44433334443336334433
3443334433633444444
```

Qwen3

LM

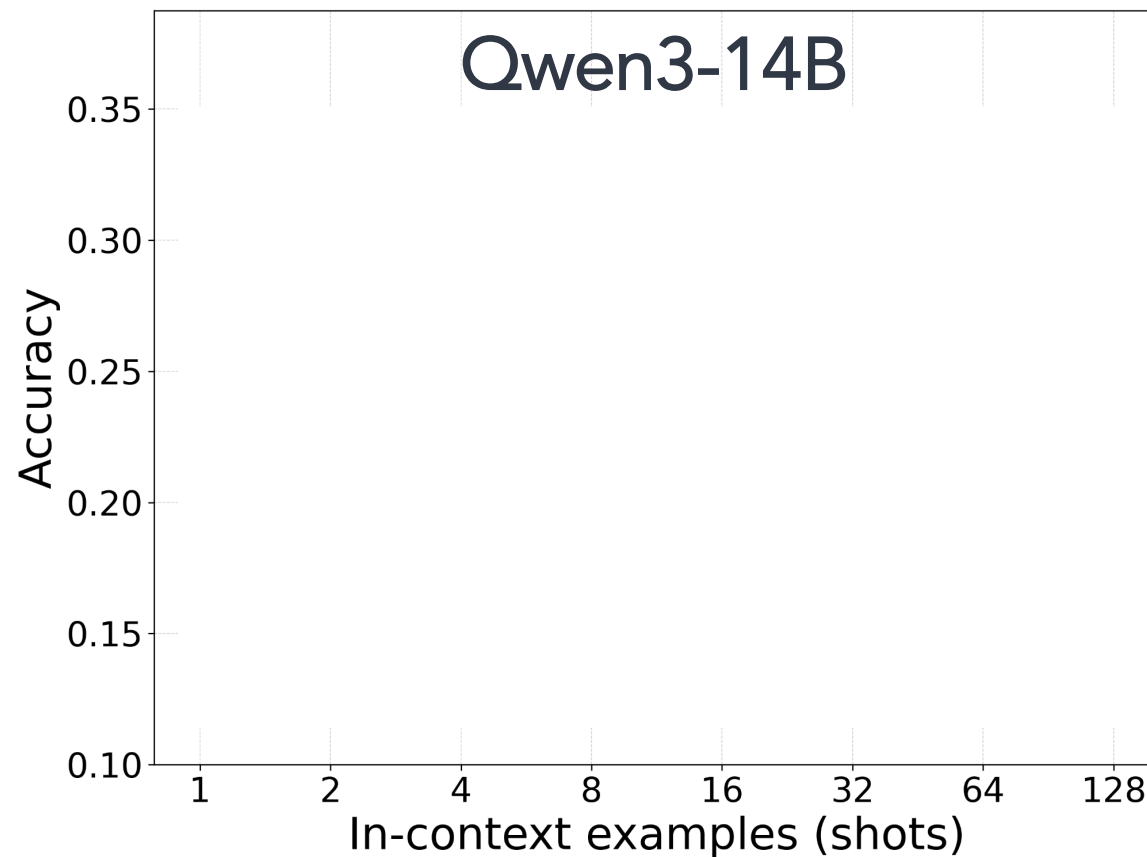
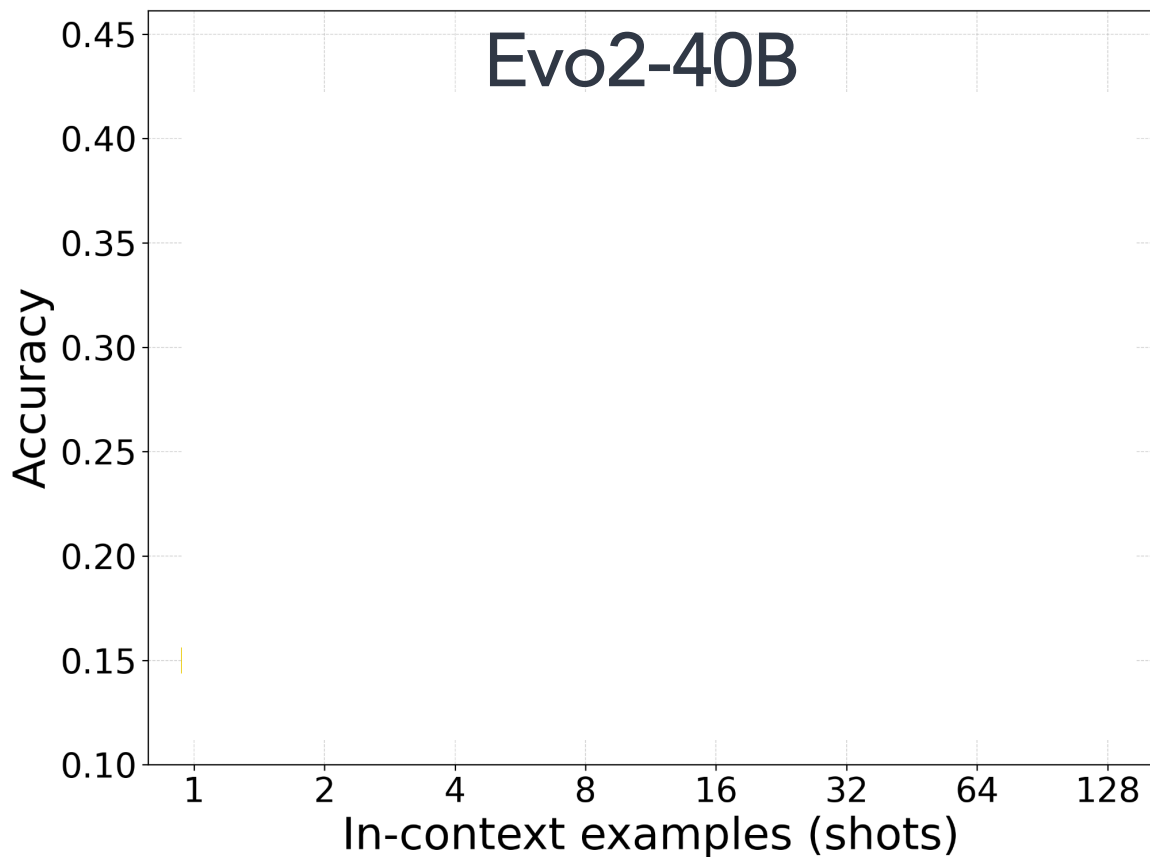
- Replace "1" with "T"
- Replace "0" with "A"
- Replace "SEP" with "G"
- Drop "->"

```
TATAAAAAAAAAAATATGTTT
AAATTTTAAATTTGTTAATT
TAATTTAATTGTTAAAAAA
```

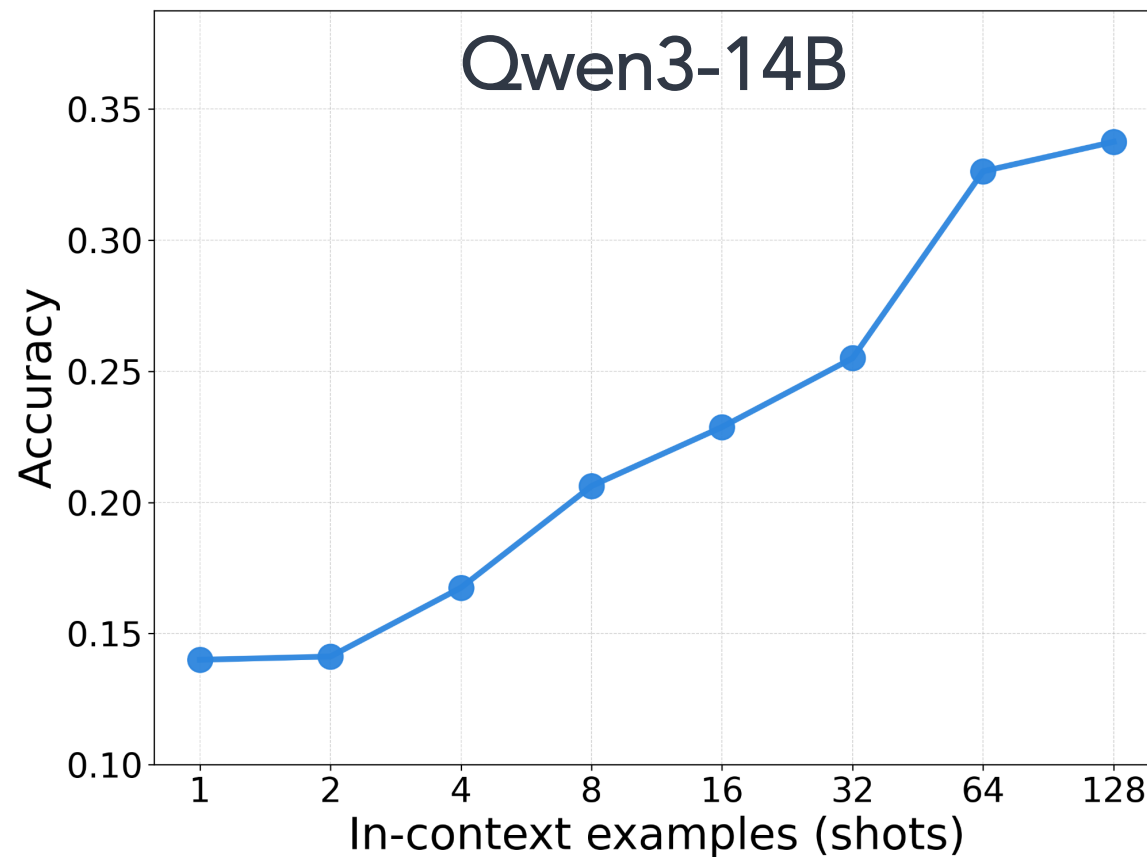
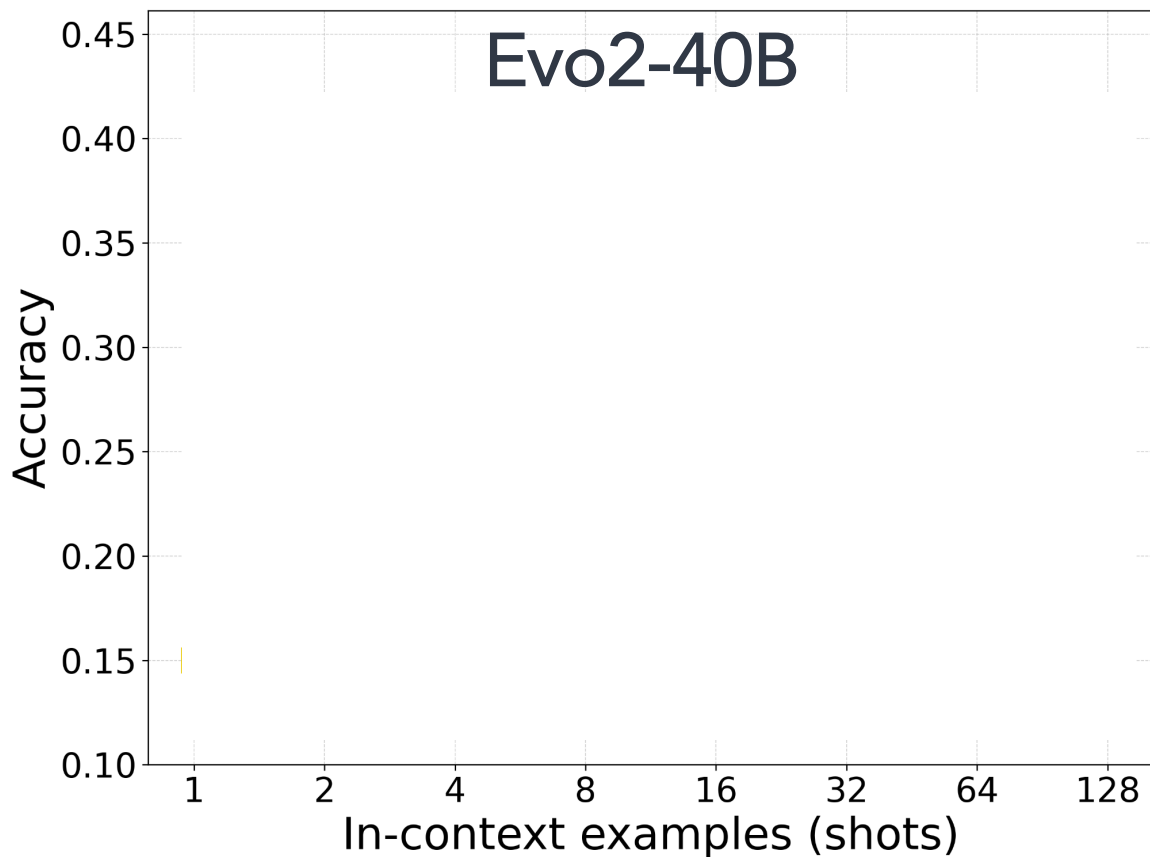
GM

Evo2

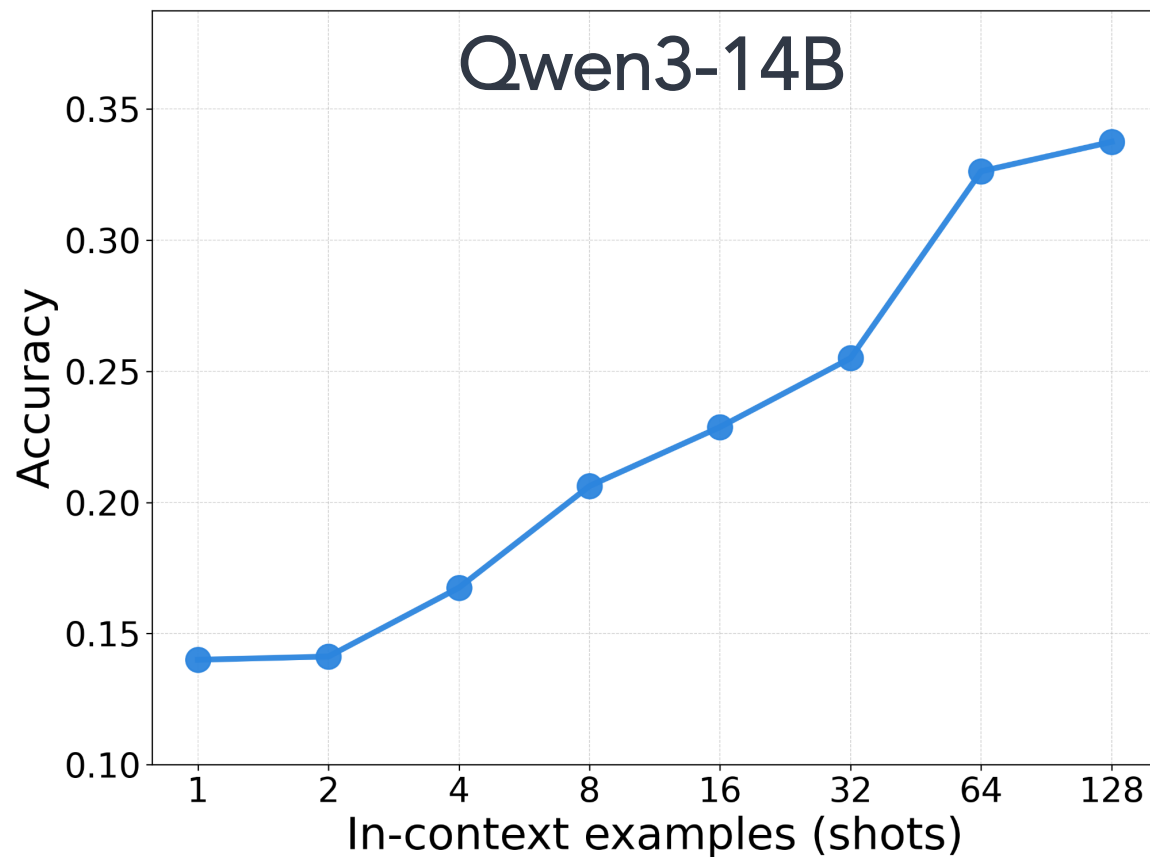
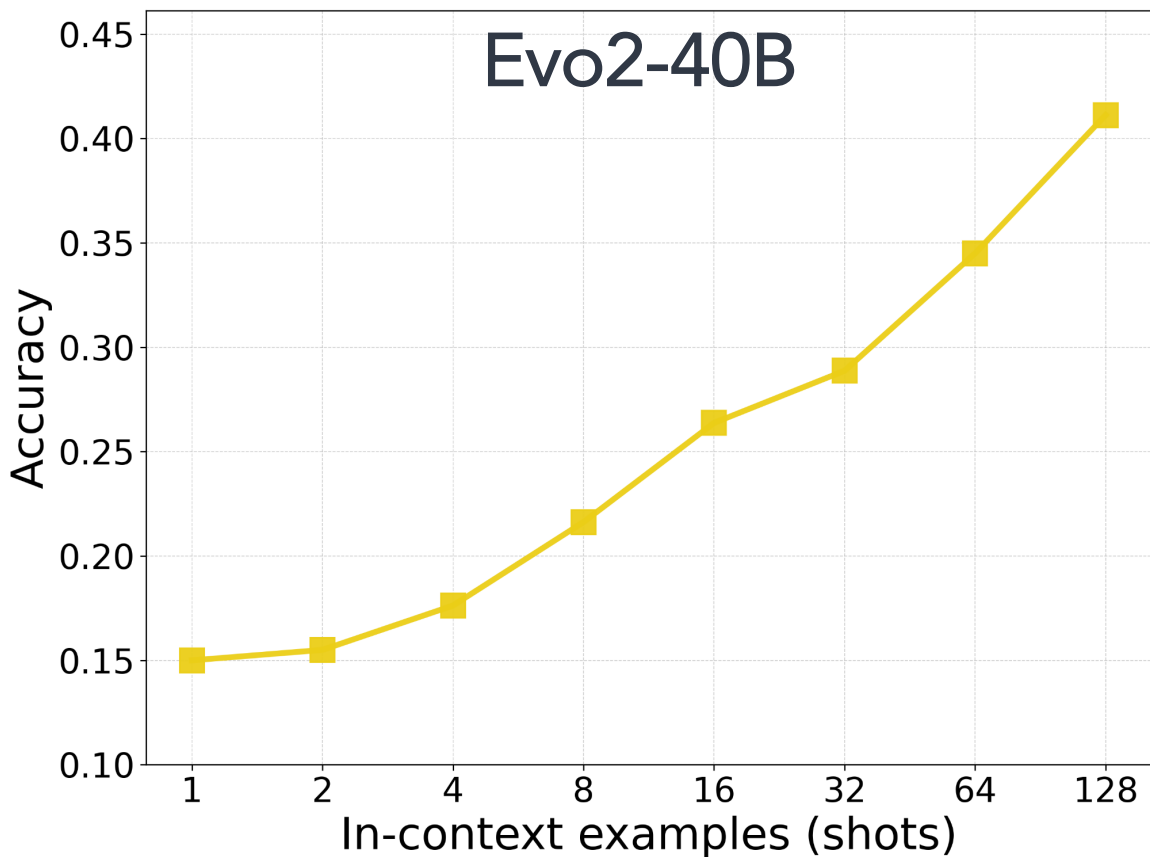
Genomic Models are In-Context Learners



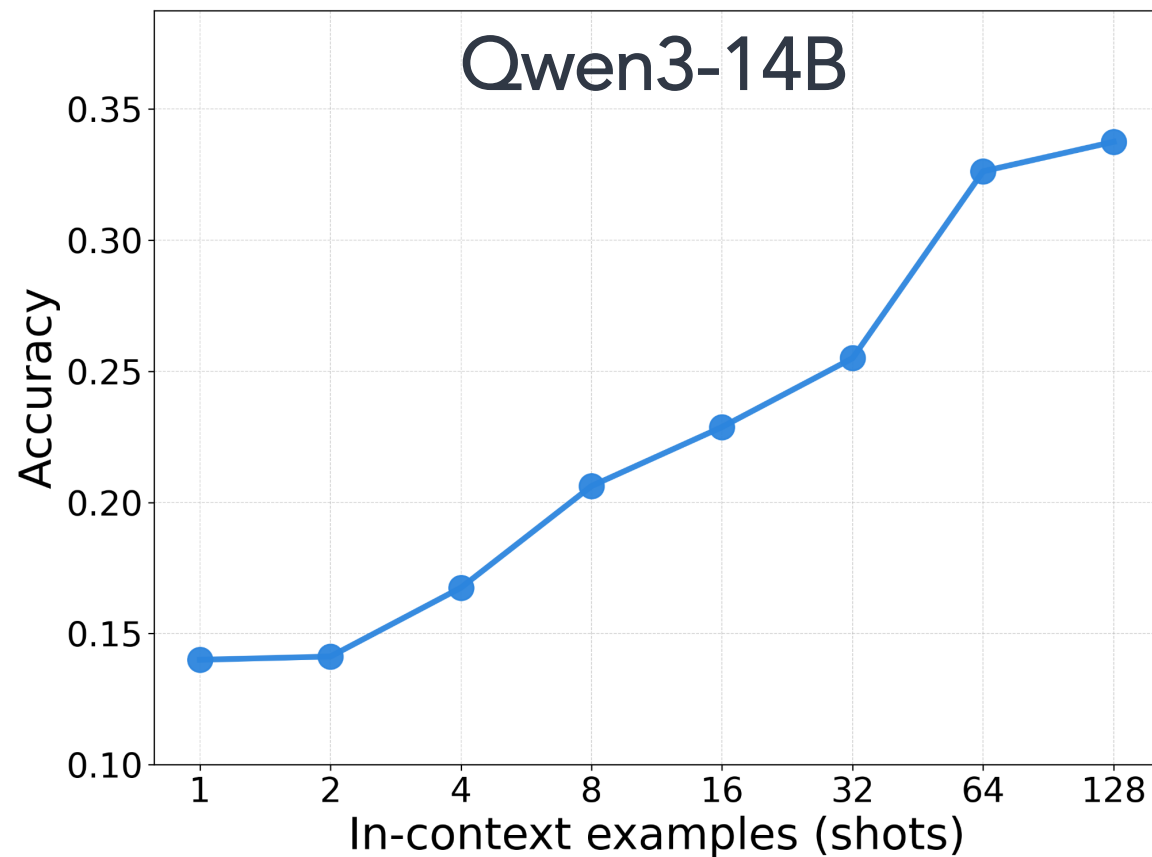
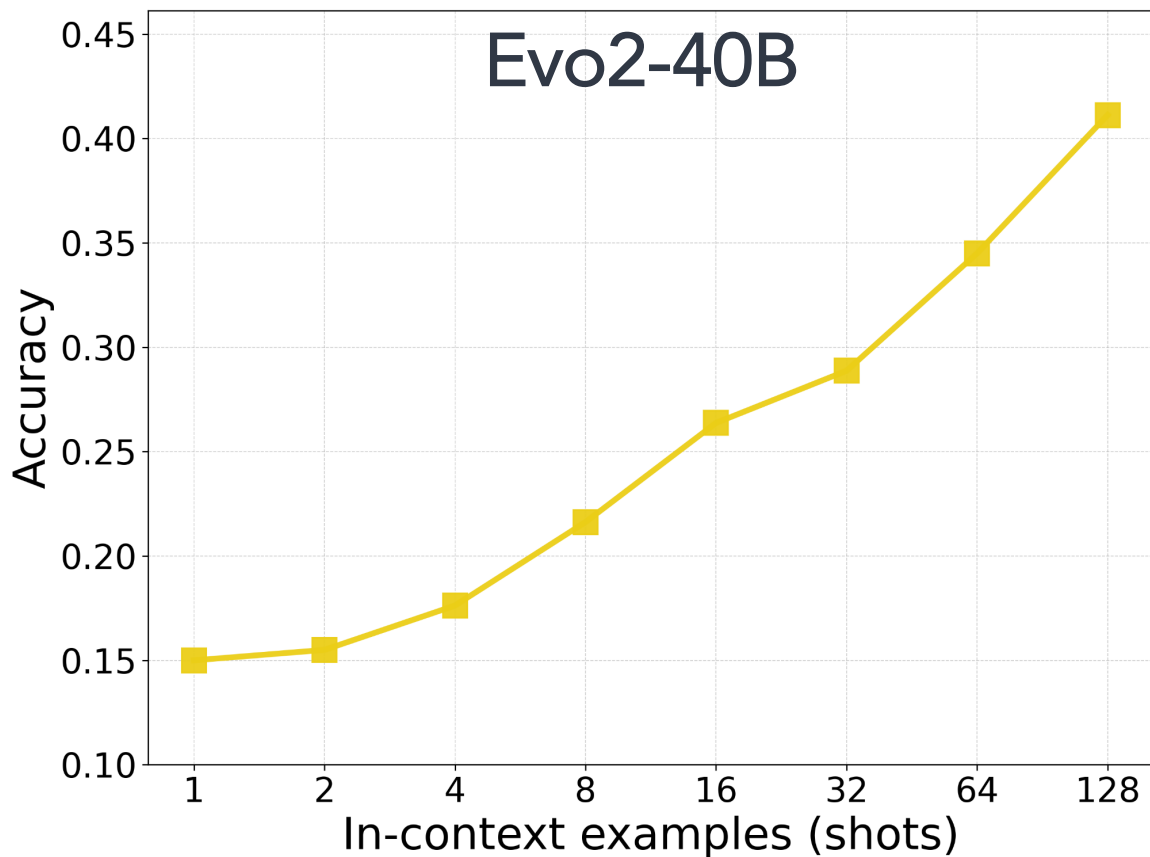
Genomic Models are In-Context Learners



Genomic Models are In-Context Learners

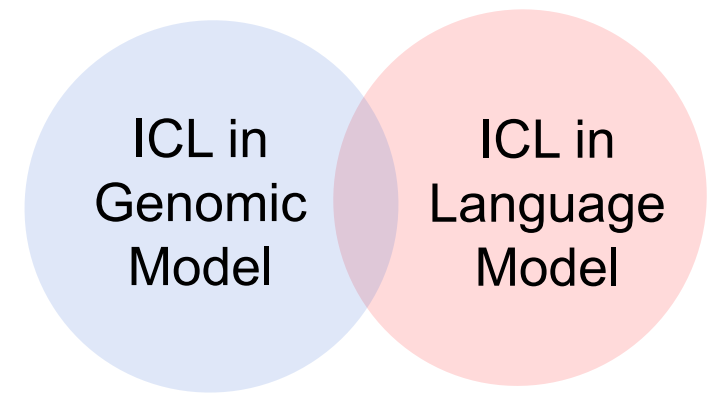


Genomic Models are In-Context Learners



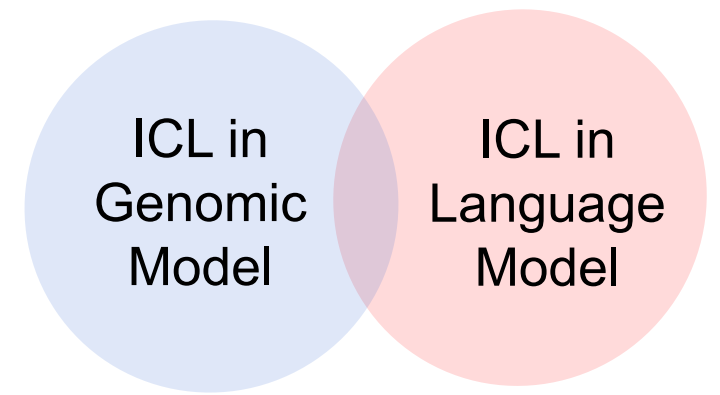
Both models exhibit **log-linear gains** in pattern induction as the number of in-context demonstrations increases.

Summary



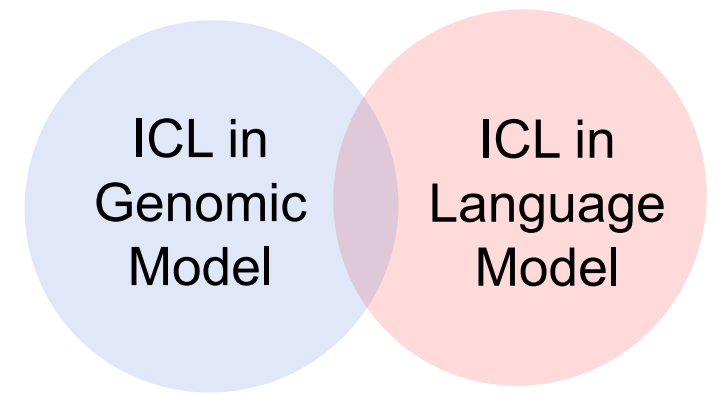
- Next-nucleotide predictors are in-context learners.
- Emergent ICL is clearly not tied to human language.
- ICL is a consequence of large autoregressive training on rich distribution.

Summary



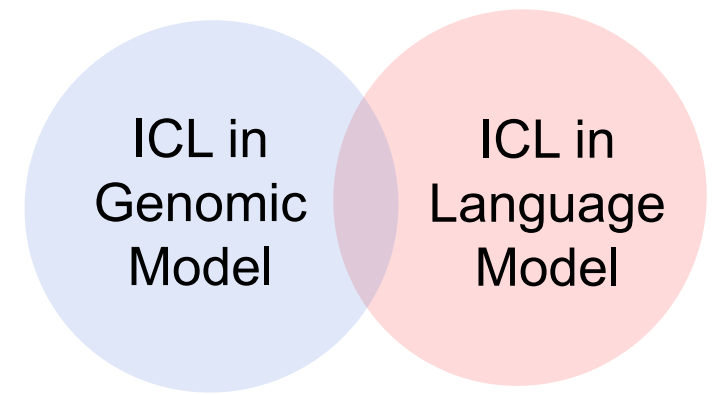
- Next-nucleotide predictors are in-context learners.
- Emergent ICL is clearly not tied to human language.
- ICL is a consequence of large autoregressive training on rich distribution.

Summary



- Next-nucleotide predictors are in-context learners.
- Emergent ICL is clearly not tied to human language.
- ICL is a consequence of large autoregressive training on rich distribution.

Summary



- Next-nucleotide predictors are in-context learners.
- Emergent ICL is clearly not tied to human language.
- ICL is a consequence of large autoregressive training on rich distribution.

Open Questions

- What other data can yield ICL? And what's common in them?
- If reasoning arises without human-interpretable tokens, what does that say about the nature of “reasoning”?
- Is it possible one can learn to reason (say in human lang) by learning from *other* modalities?

Open Questions

- What other data can yield ICL? And what's common in them?
- If reasoning arises without human-interpretable tokens, what does that say about the nature of “reasoning”?
- Is it possible one can learn to reason (say in human lang) by learning from *other* modalities?

Open Questions

- What other data can yield ICL? And what's common in them?

✓ Human language

Today I'm presenting ...



- If reasoning arises without human-interpretable tokens, what does that say about the nature of “reasoning”?
- Is it possible one can learn to reason (say in human lang) by learning from *other* modalities?

Open Questions

- What other data can yield ICL? And what's common in them?

✓ Human language

Today I'm presenting ...

✓ DNA sequences

T A T A A A A A A T A T G T T T ...



- If reasoning arises without human-interpretable tokens, what does that say about the nature of “reasoning”?
- Is it possible one can learn to reason (say in human lang) by learning from *other* modalities?

Open Questions

- What other data can yield ICL? And what's common in them?

✓ Human language

Today I'm presenting ...

✓ DNA sequences

T A T A A A A A A T A T G T T T ...

? Brain signals

12, 15, 18, 14, 10, 8, 11, 17, 22, ...



- If reasoning arises without human-interpretable tokens, what does that say about the nature of “reasoning”?
- Is it possible one can learn to reason (say in human lang) by learning from *other* modalities?

Open Questions

- What other data can yield ICL? And what's common in them?

✓ Human language

Today I'm presenting ...

✓ DNA sequences

T A T A A A A A A T A T G T T T ...

? Brain signals

12, 15, 18, 14, 10, 8, 11, 17, 22, ...

? Time series

18934, 18951, 18920, 18885, ...



- If reasoning arises without human-interpretable tokens, what does that say about the nature of “reasoning”?
- Is it possible one can learn to reason (say in human lang) by learning from *other* modalities?

Open Questions

- What other data can yield ICL? And what's common in them?

✓ Human language

Today I'm presenting ...

✓ DNA sequences

T A T A A A A A A T A T G T T T ...

? Brain signals

12, 15, 18, 14, 10, 8, 11, 17, 22, ...

? Time series

18934, 18951, 18920, 18885, ...

? Random sequences

plorv amk xe ##tion 3r nalf qu ty ...



- If reasoning arises without human-interpretable tokens, what does that say about the nature of “reasoning”?
- Is it possible one can learn to reason (say in human lang) by learning from *other* modalities?

Open Questions

- What other data can yield ICL? And what's common in them?

✓ Human language

Today I'm presenting ...

✓ DNA sequences

T A T A A A A A A T A T G T T T ...

? Brain signals

12, 15, 18, 14, 10, 8, 11, 17, 22, ...

? Time series

18934, 18951, 18920, 18885, ...

? Random sequences

plorv amk xe ##tion 3r nalf qu ty ...



- If reasoning arises without human-interpretable tokens, what does that say about the nature of “reasoning”?
- Is it possible one can learn to reason (say in human lang) by learning from *other* modalities?

Open Questions

- What other data can yield ICL? And what's common in them?

✓ Human language

Today I'm presenting ...

✓ DNA sequences

T A T A A A A A A T A T G T T T ...

? Brain signals

12, 15, 18, 14, 10, 8, 11, 17, 22, ...

? Time series

18934, 18951, 18920, 18885, ...

? Random sequences

plorv amk xe ##tion 3r nalf qu ty ...



- If reasoning arises without human-interpretable tokens, what does that say about the nature of “reasoning”?
- Is it possible one can learn to reason (say in human lang) by learning from *other* modalities?

Today: Rethinking the Reasoning Stack



1. How does in-context learning (ICL) **emerge**? (pre-training)



2. How do we **distill ICL**? (pre \rightarrow post)

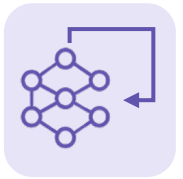


3. How do we maintain **diversity**? (post-training)

Today: Rethinking the Reasoning Stack



1. How does in-context learning (ICL) **emerge**? (pre-training)



2. How do we **distill ICL**? (pre → post)



3. How do we maintain **diversity**? (post-training)

Adaptation: Problem Setup

Adaptation: Problem Setup

A yellow rounded square with a thin orange border containing the letters "LM" in a bold, dark blue font.

Adaptation: Problem Setup

task $\mathcal{T} : \mathcal{X} \rightarrow \mathcal{Y}$

N labeled data $\left\{ \begin{array}{l} X_1 \rightarrow Y_1 \\ \vdots \\ X_N \rightarrow Y_N \end{array} \right.$

LM

Adaptation: Problem Setup

- Given a **pre-trained LM**, we want to adapt it to our input-output pairs.

task $\mathcal{T} : \mathcal{X} \rightarrow \mathcal{Y}$

N labeled data $\left\{ \begin{array}{l} X_1 \rightarrow Y_1 \\ \vdots \\ X_N \rightarrow Y_N \end{array} \right.$

LM

- Various use-cases: (1) model alignment; (2) model specialization.

Two Adaptation Paradigms

Supervised Fine-tuning (SFT)

Two Adaptation Paradigms

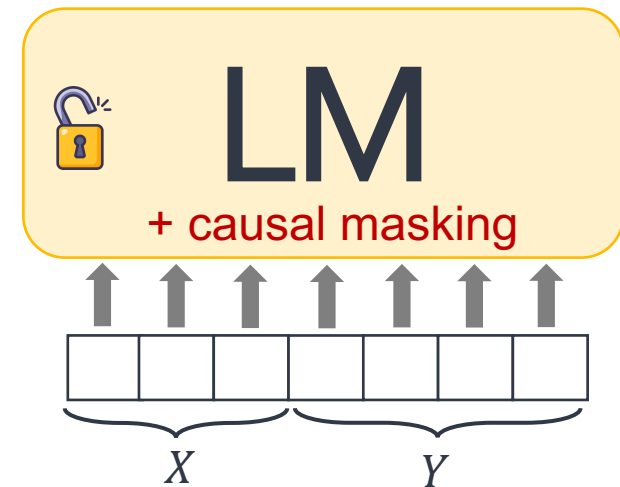
Supervised Fine-tuning (SFT)



Training time

Two Adaptation Paradigms

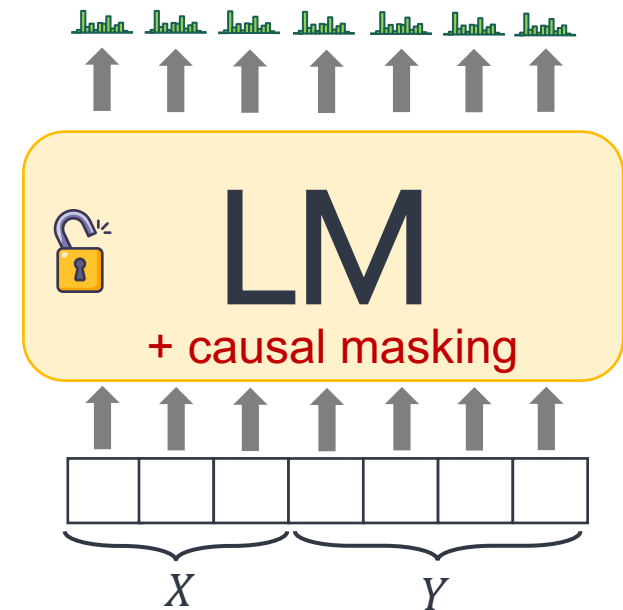
Supervised Fine-tuning (SFT)



Training time

Two Adaptation Paradigms

Supervised Fine-tuning (SFT)

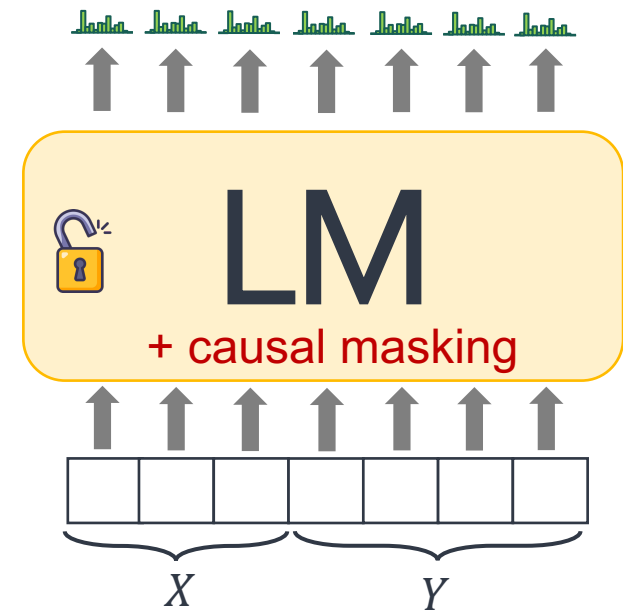


Training time

Two Adaptation Paradigms

Supervised Fine-tuning (SFT)

$$\text{CE loss: } \ell_t = -\log P(y_t | y_{<t}, X)$$



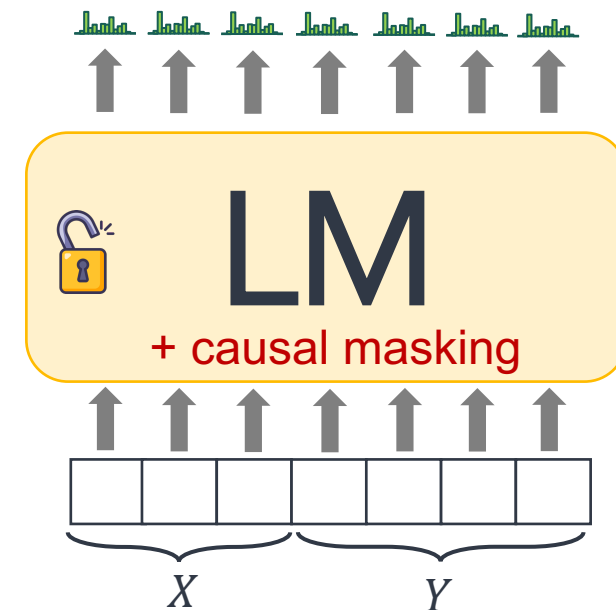
Training time

Two Adaptation Paradigms

Supervised Fine-tuning (SFT)

$$\text{CE loss: } \ell_t = -\log P(y_t | y_{<t}, X)$$

$$\ell_1 + \ell_2 + \ell_3 + \ell_4 = \mathcal{L}$$

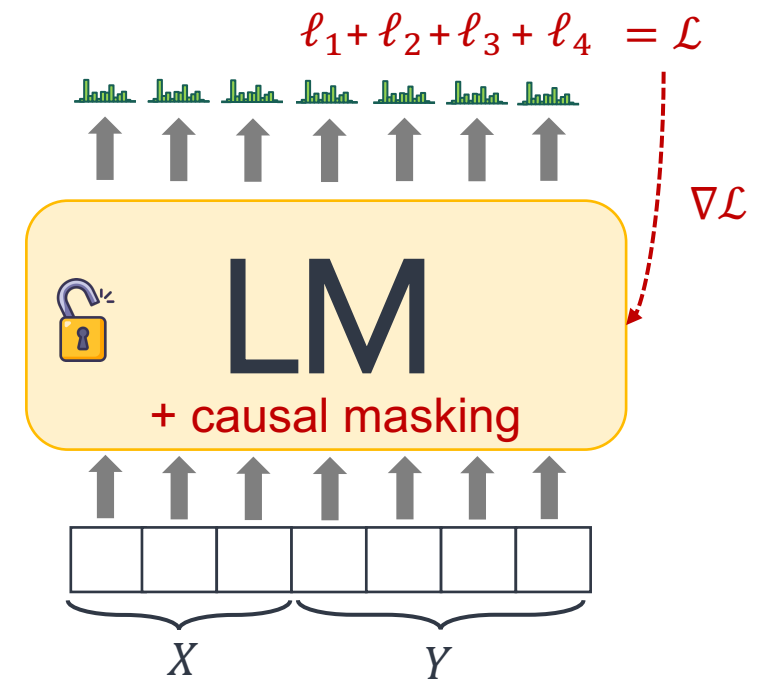


Training time

Two Adaptation Paradigms

Supervised Fine-tuning (SFT)

$$\text{CE loss: } \ell_t = -\log P(y_t | y_{<t}, X)$$



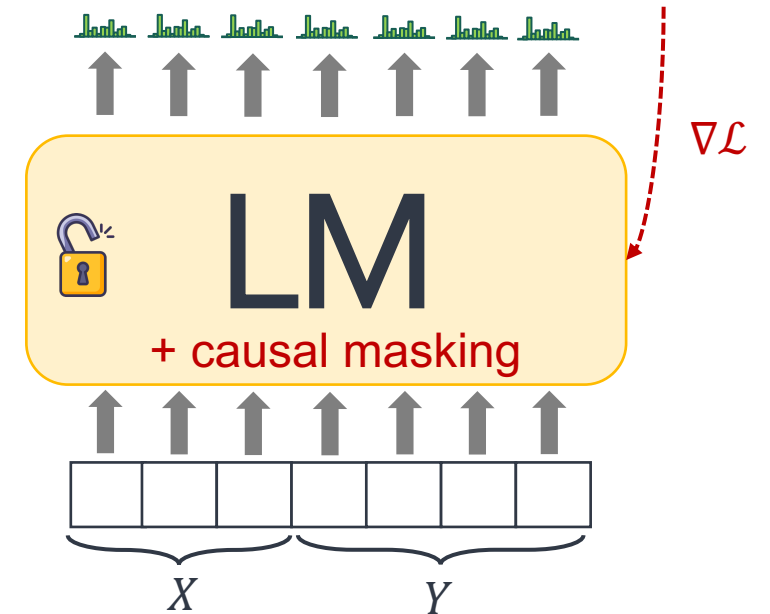
Training time

Two Adaptation Paradigms

Supervised Fine-tuning (SFT)

$$\text{CE loss: } \ell_t = -\log P(y_t | y_{<t}, X)$$

$$\ell_1 + \ell_2 + \ell_3 + \ell_4 = \mathcal{L}$$



Training time

Two Adaptation Paradigms

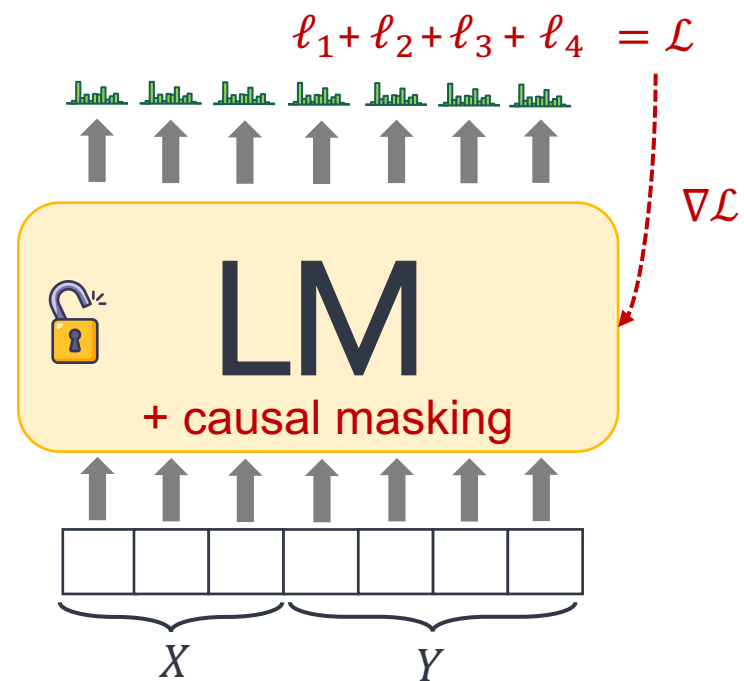
Supervised Fine-tuning (SFT)



Test time



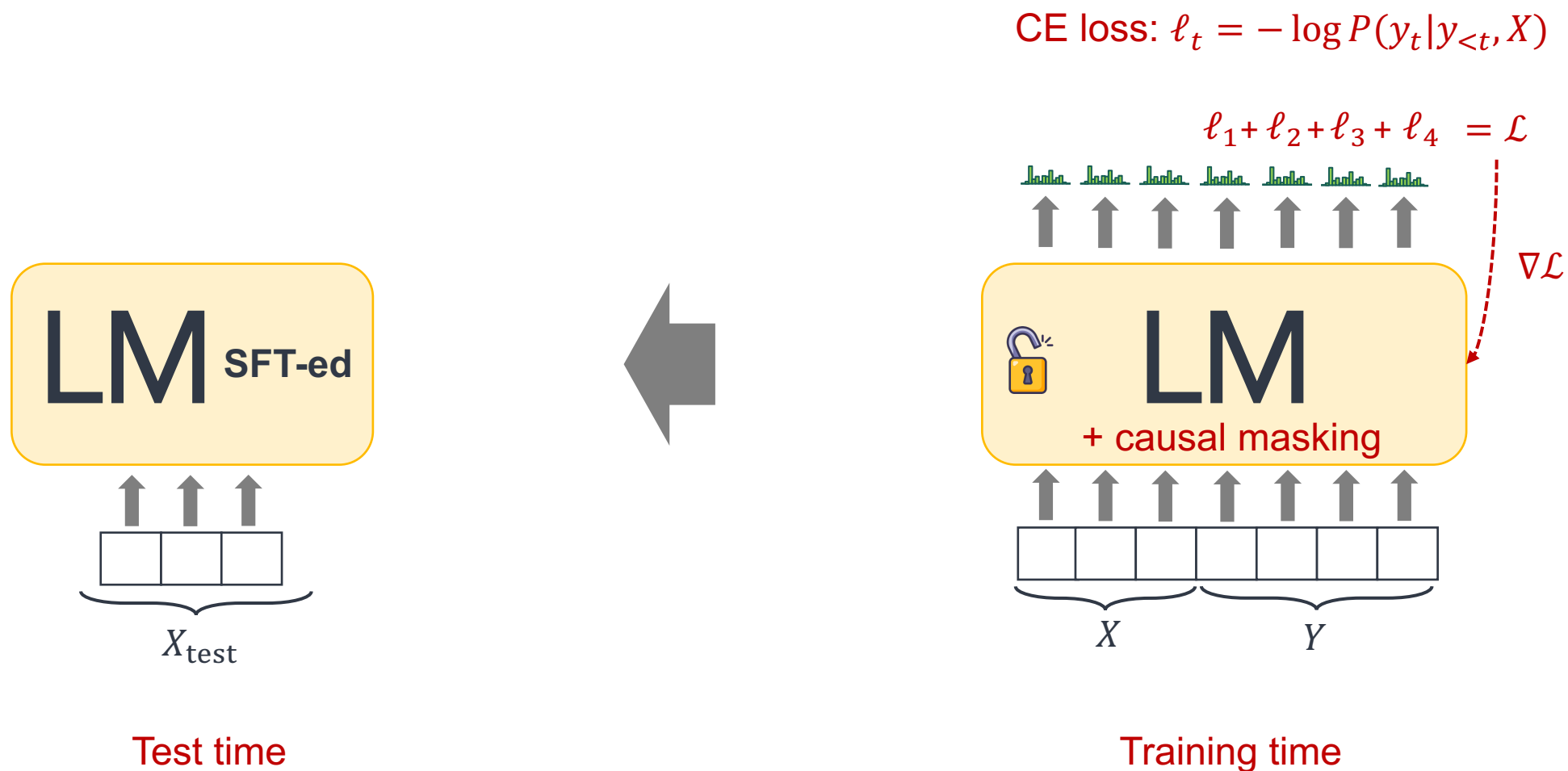
$$\text{CE loss: } \ell_t = -\log P(y_t | y_{<t}, X)$$



Training time

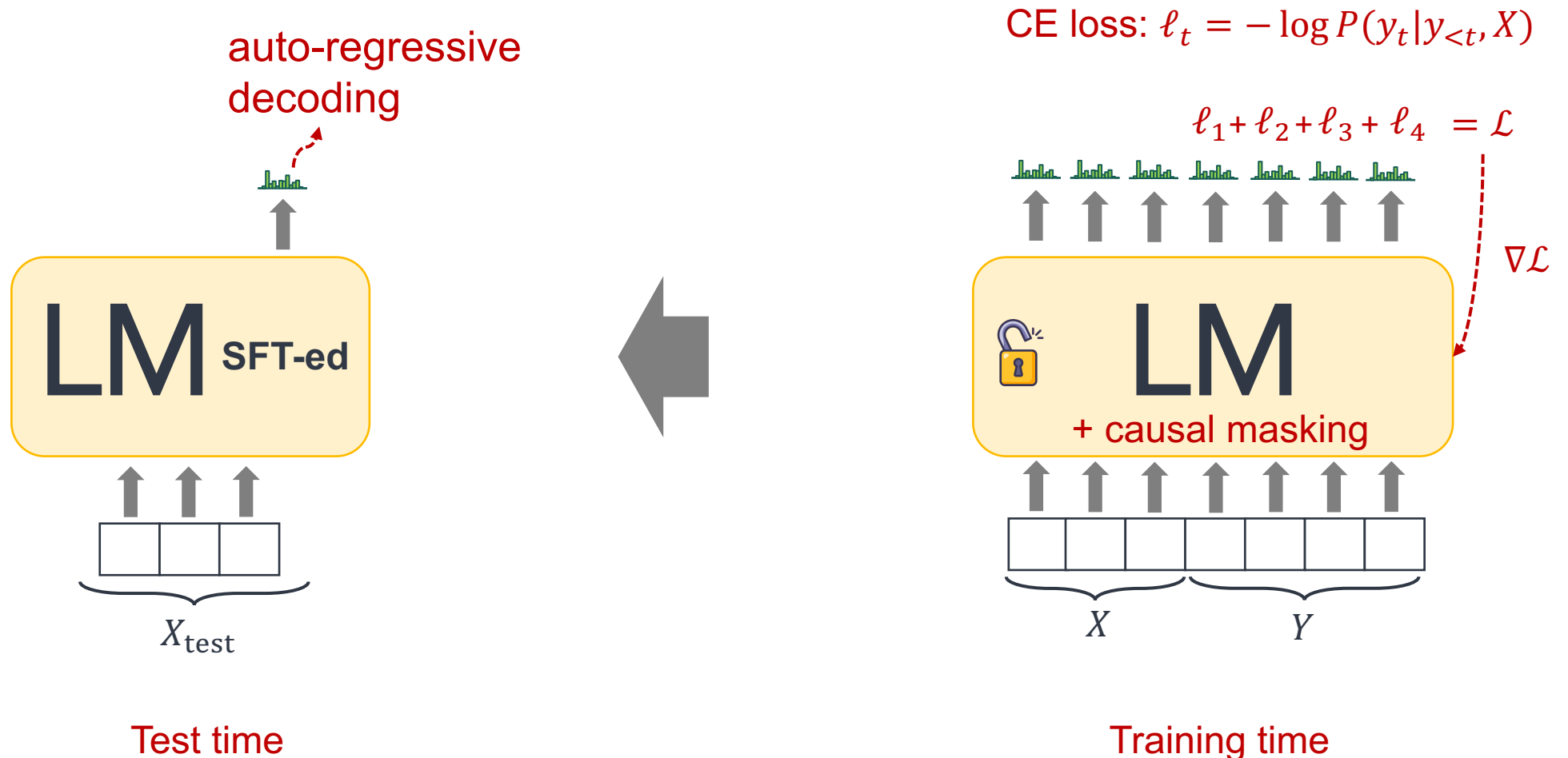
Two Adaptation Paradigms

Supervised Fine-tuning (SFT)



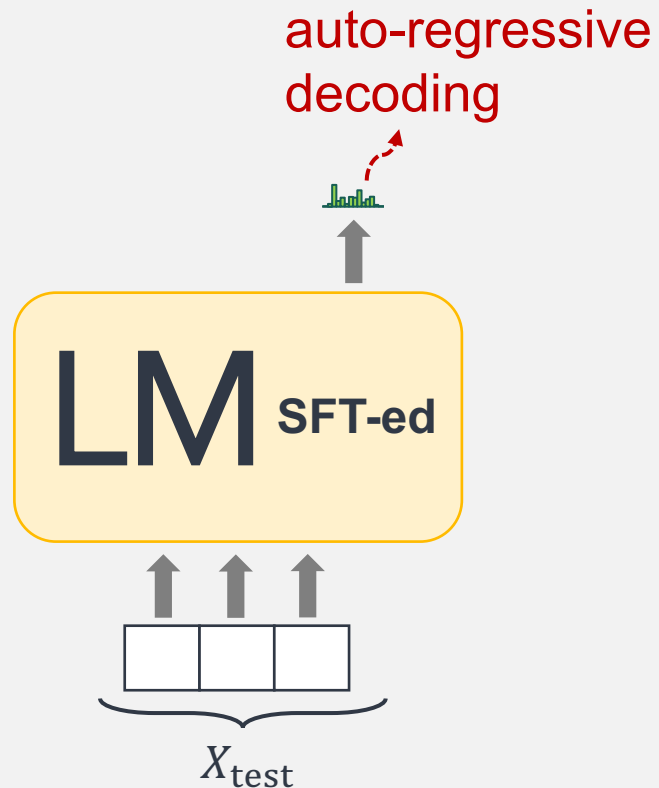
Two Adaptation Paradigms

Supervised Fine-tuning (SFT)



Two Adaptation Paradigms

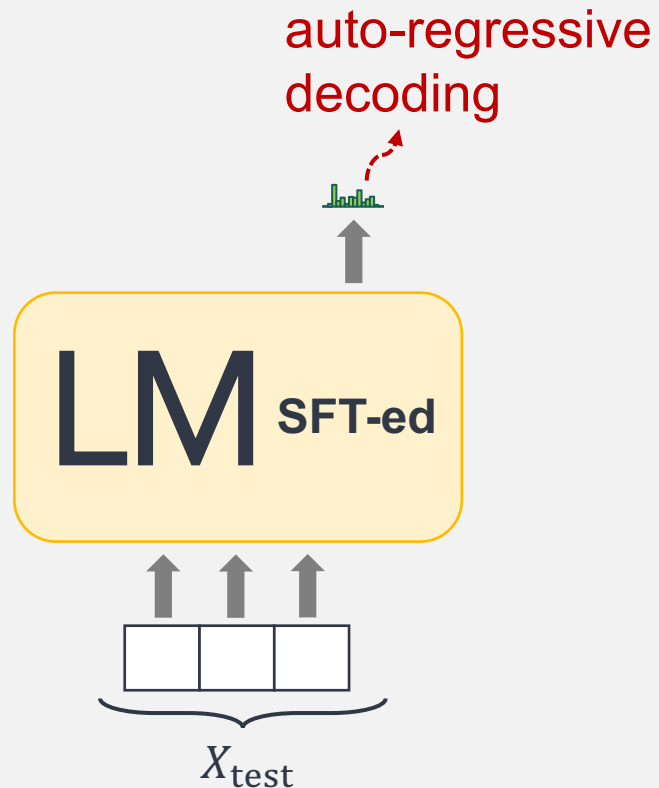
Supervised Fine-tuning (SFT)



Test time

Two Adaptation Paradigms

Supervised Fine-tuning (SFT)



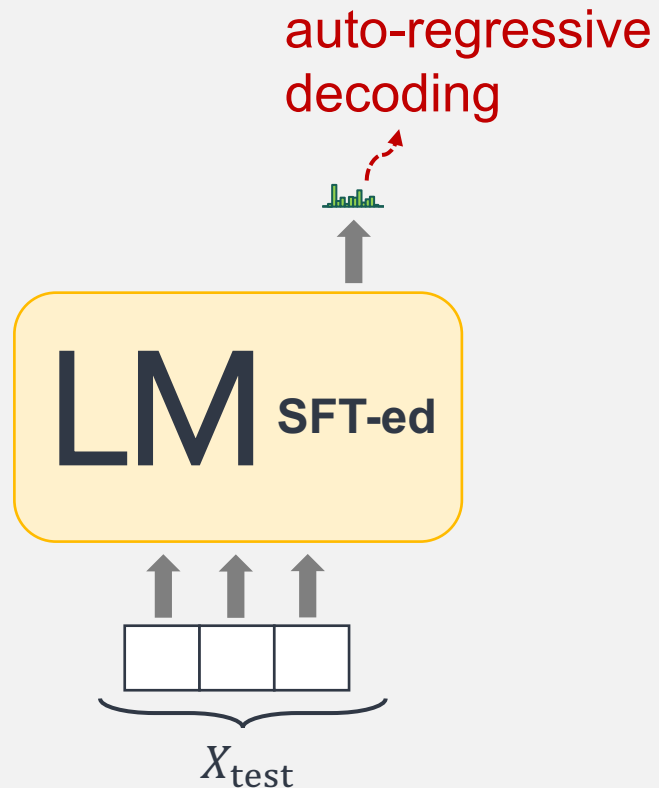
Test time

In-Context Learning (ICL)

Test time

Two Adaptation Paradigms

Supervised Fine-tuning (SFT)



Test time

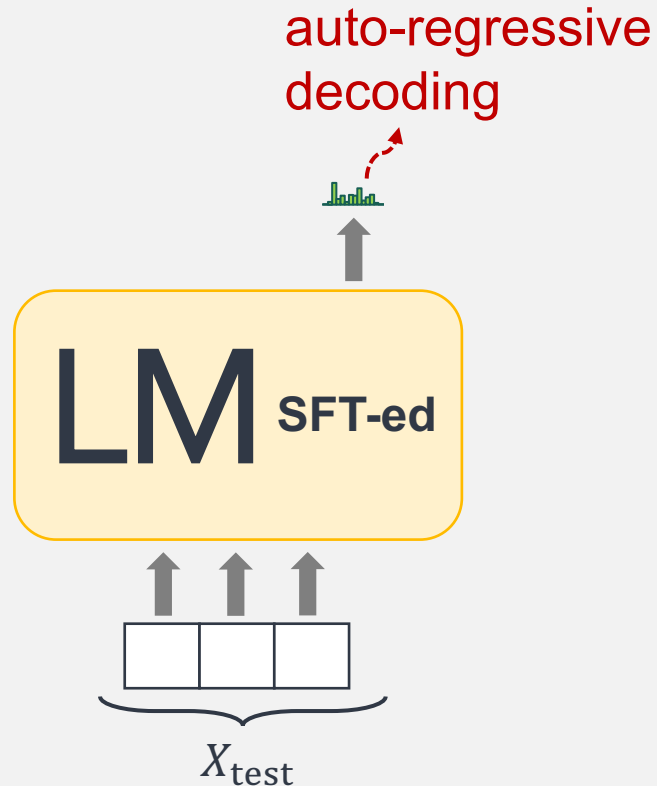
In-Context Learning (ICL)



Test time

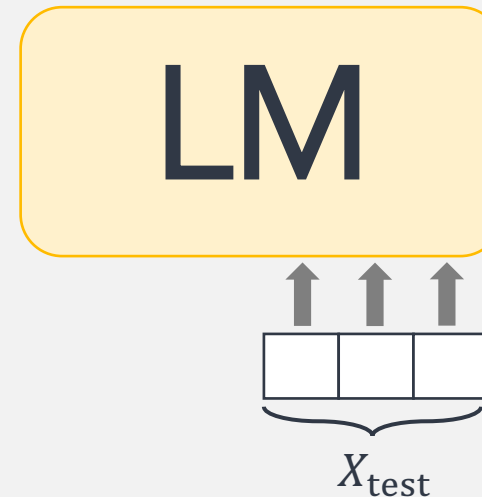
Two Adaptation Paradigms

Supervised Fine-tuning (SFT)



Test time

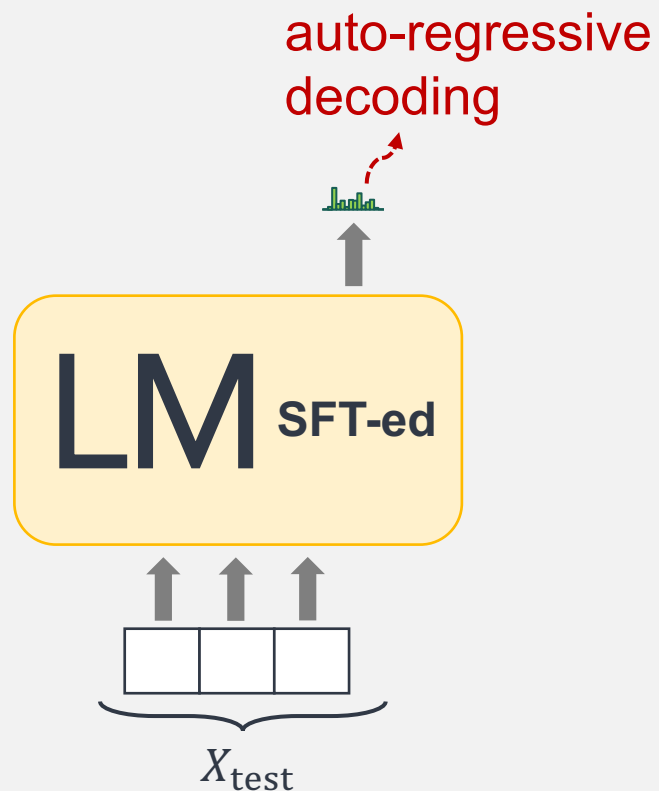
In-Context Learning (ICL)



Test time

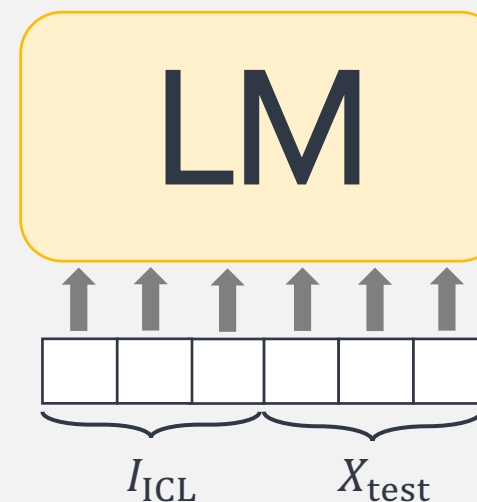
Two Adaptation Paradigms

Supervised Fine-tuning (SFT)



Test time

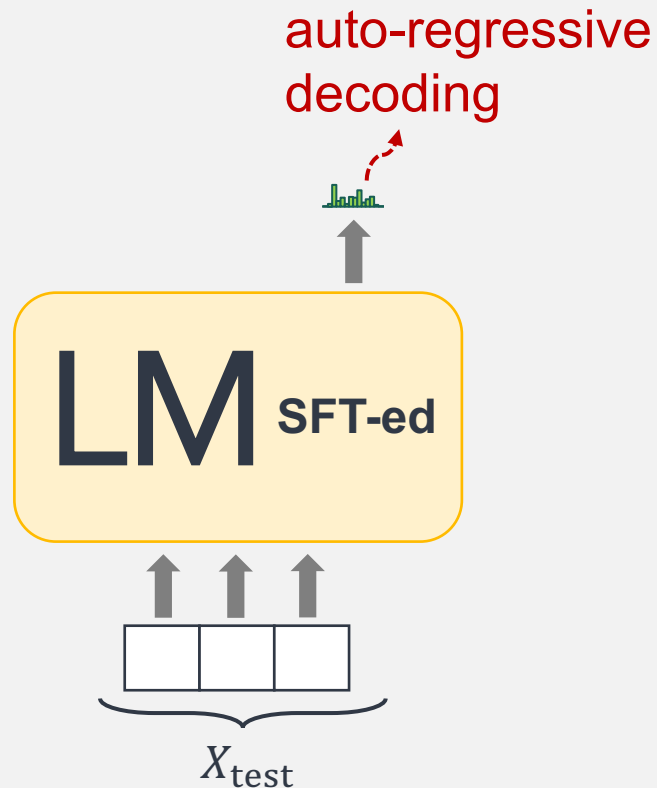
In-Context Learning (ICL)



Test time

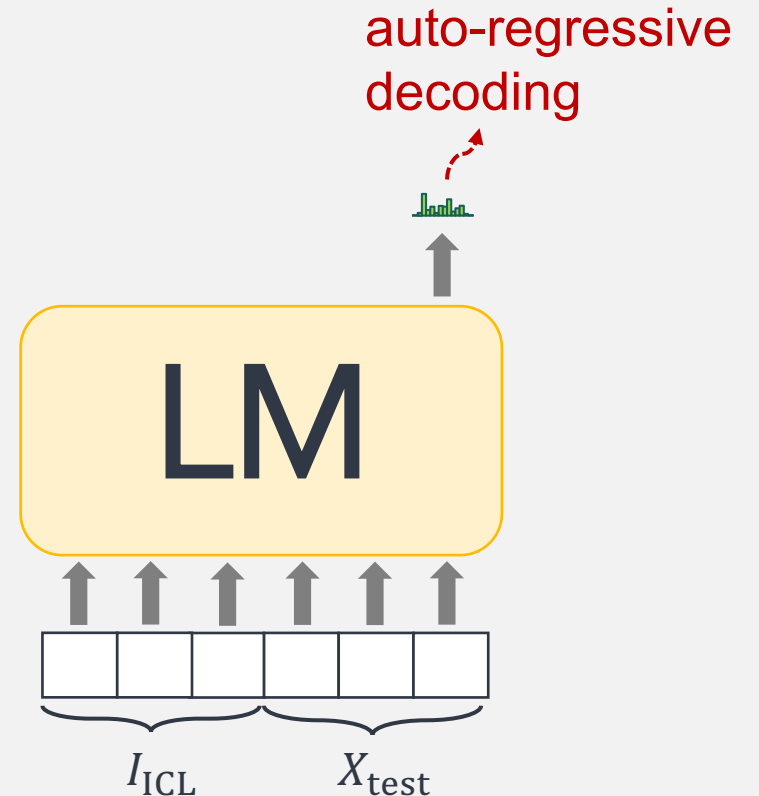
Two Adaptation Paradigms

Supervised Fine-tuning (SFT)



Test time

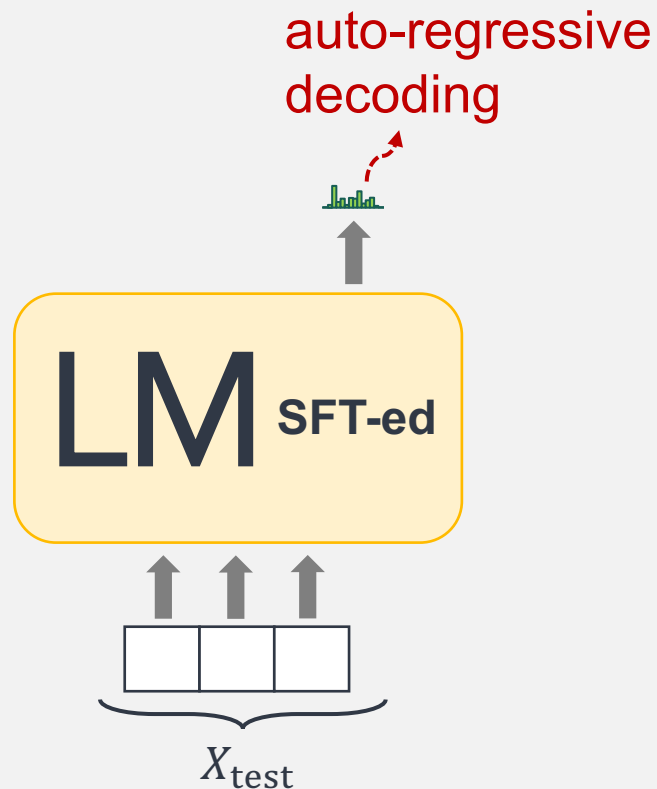
In-Context Learning (ICL)



Test time

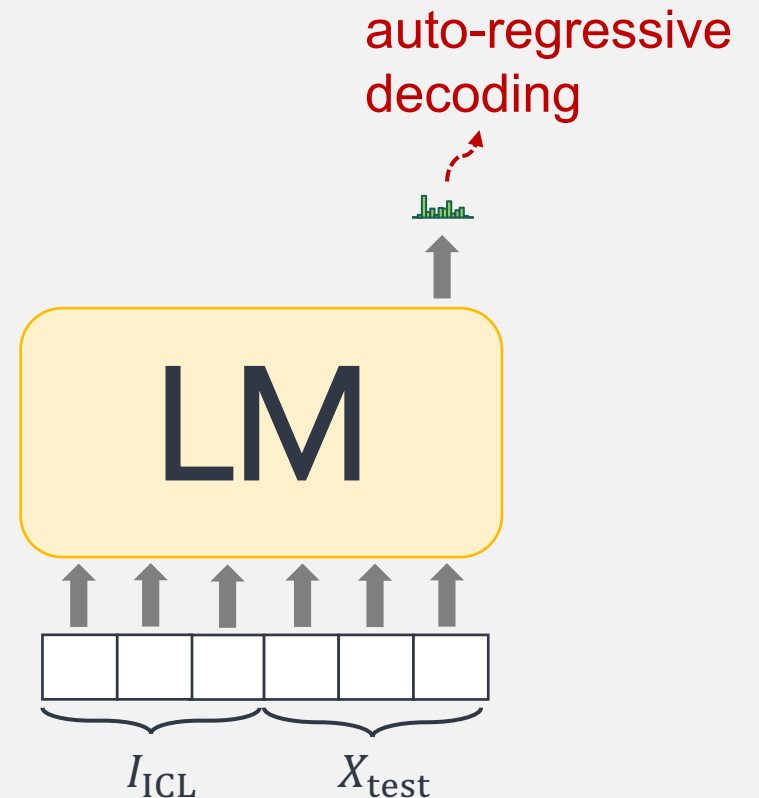
Which one would you use?

Supervised Fine-tuning (SFT)



Test time

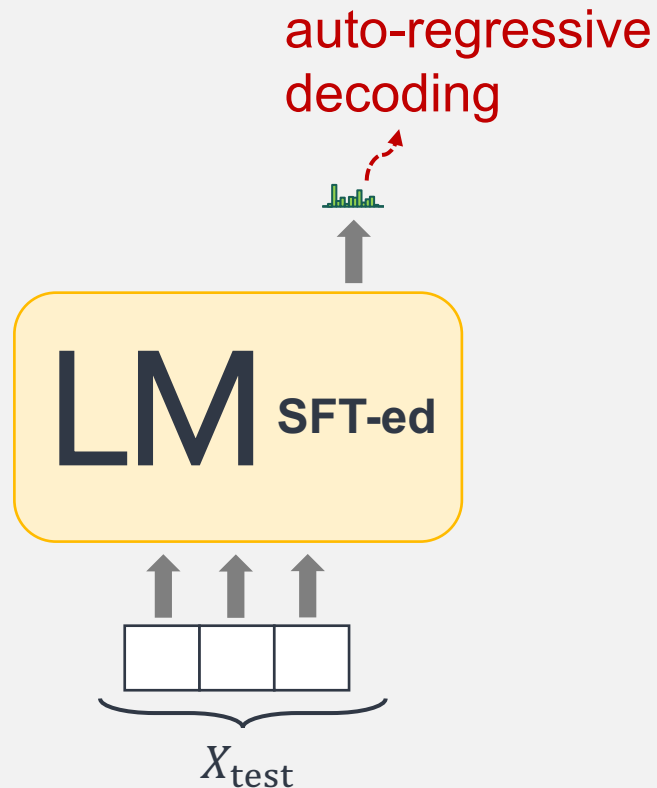
In-Context Learning (ICL)



Test time

Which one would you use?

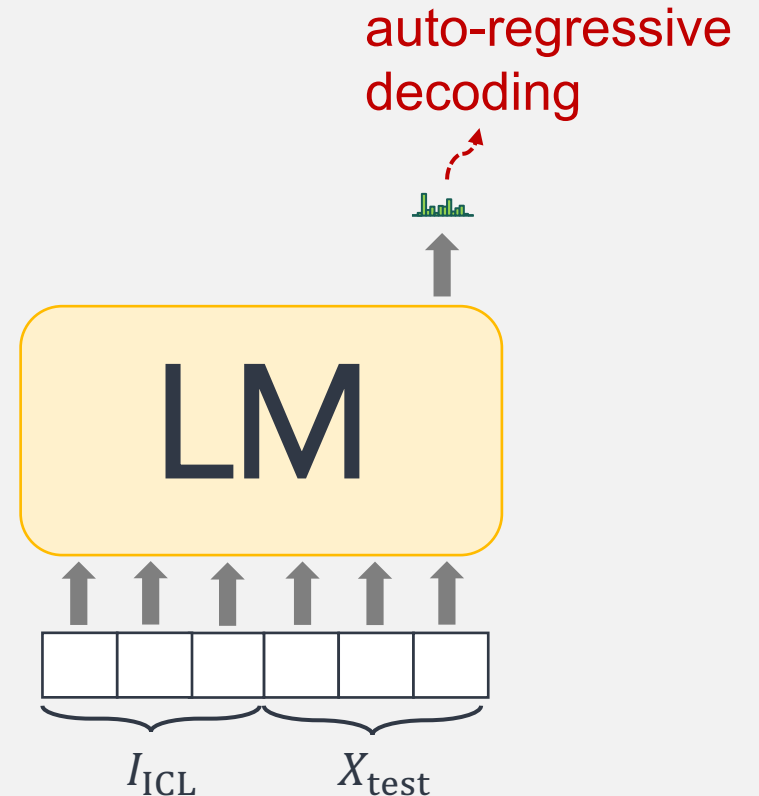
Supervised Fine-tuning (SFT)



Test time



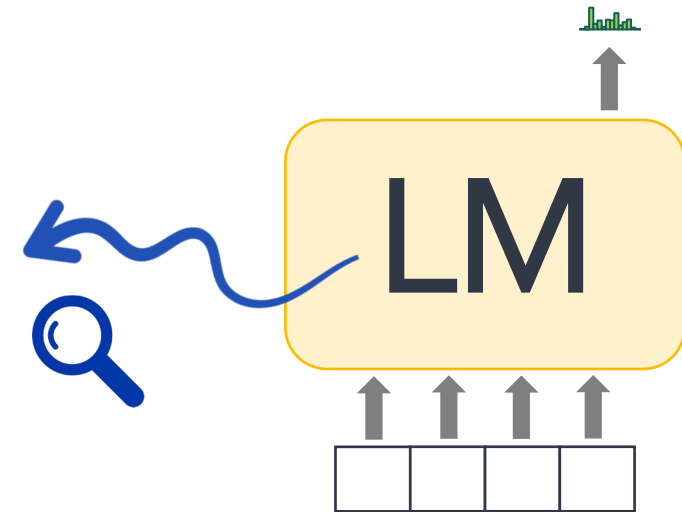
In-Context Learning (ICL)



Test time

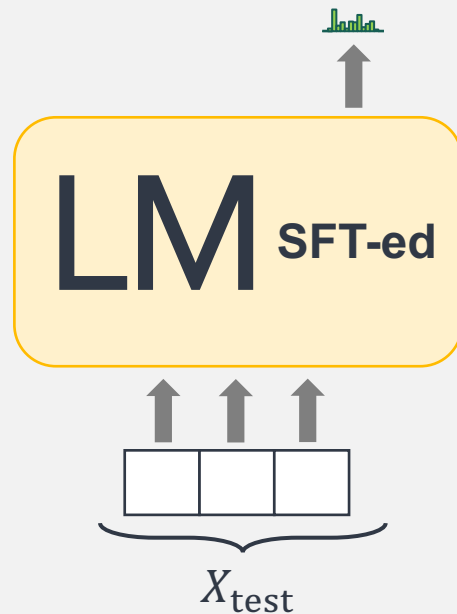
How does ICL compare to SFT?

Activations at the output
of Self-Attention Module

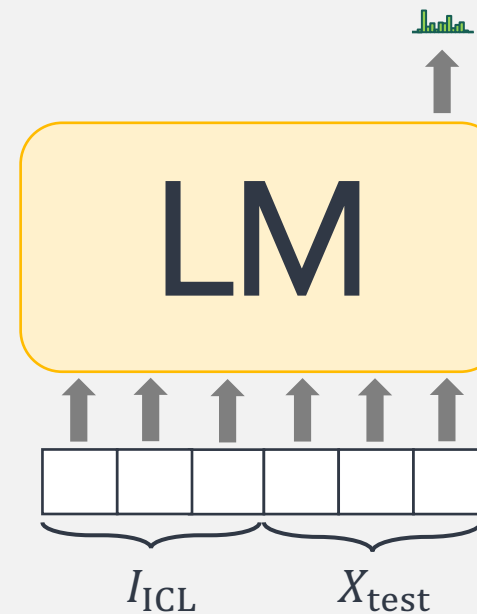


How does ICL compare to SFT?

Supervised Fine-tuning (SFT)

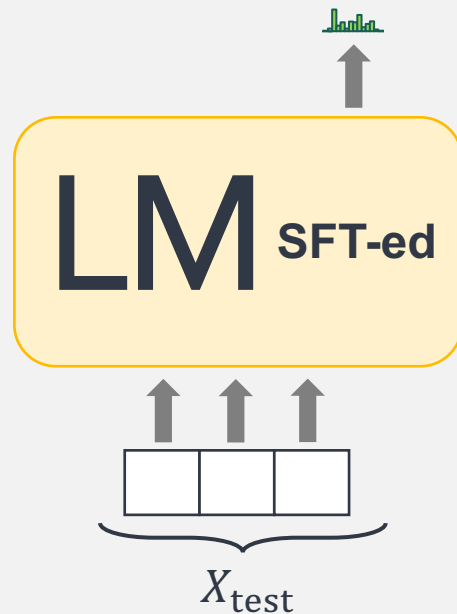


In-Context Learning (ICL)

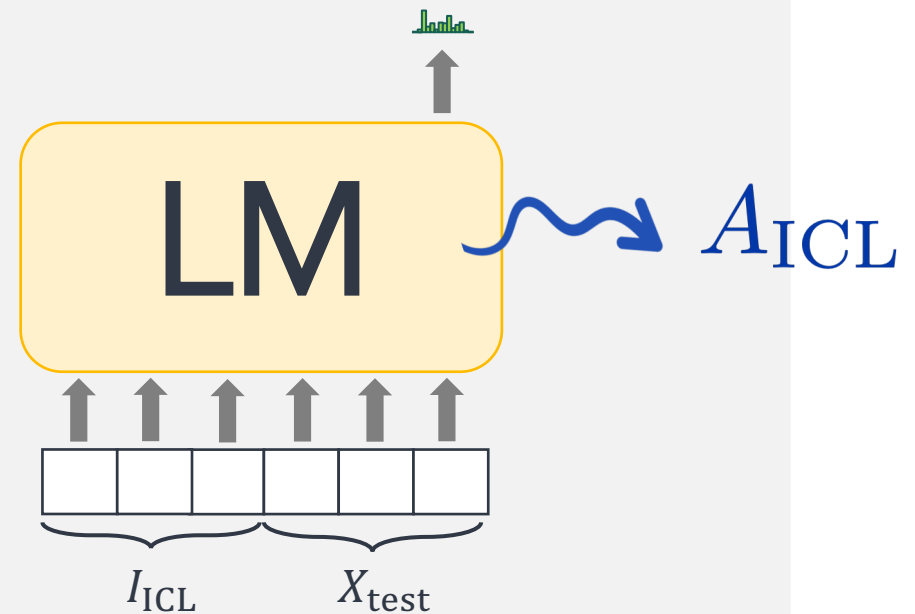


How does ICL compare to SFT?

Supervised Fine-tuning (SFT)

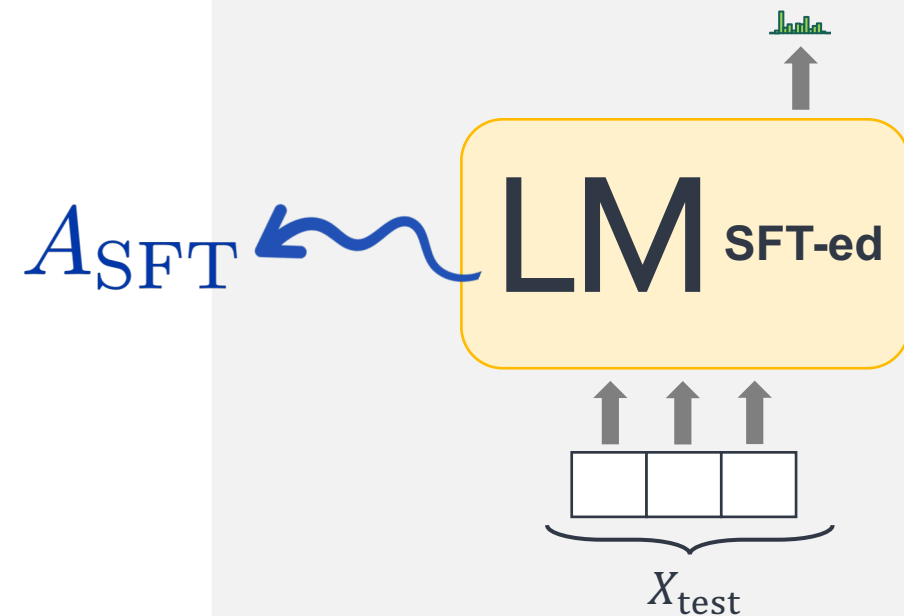


In-Context Learning (ICL)

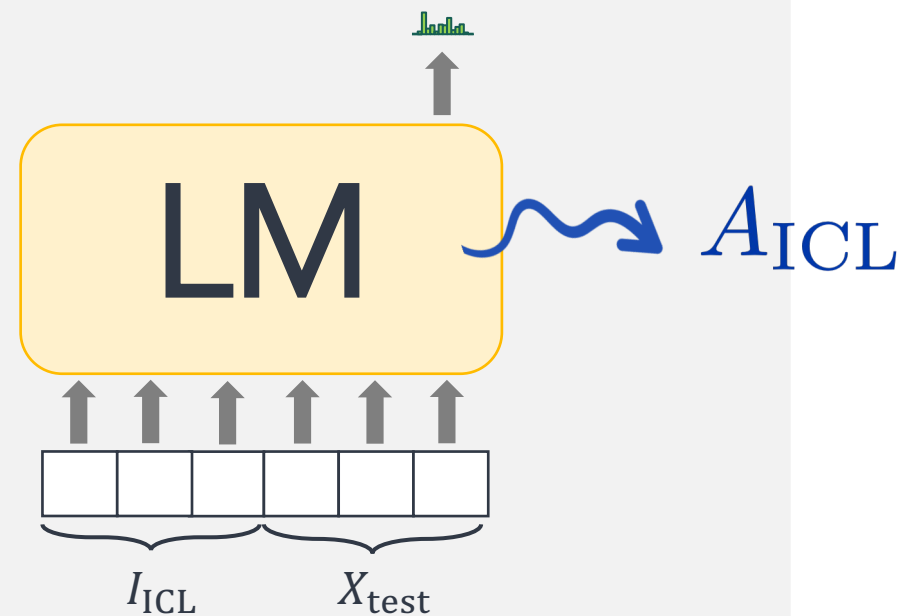


How does ICL compare to SFT?

Supervised Fine-tuning (SFT)

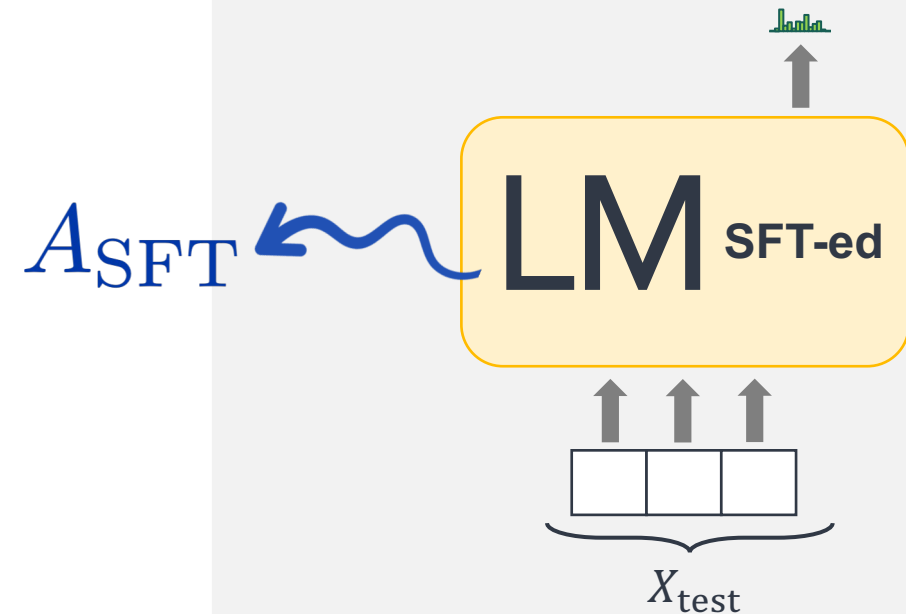


In-Context Learning (ICL)

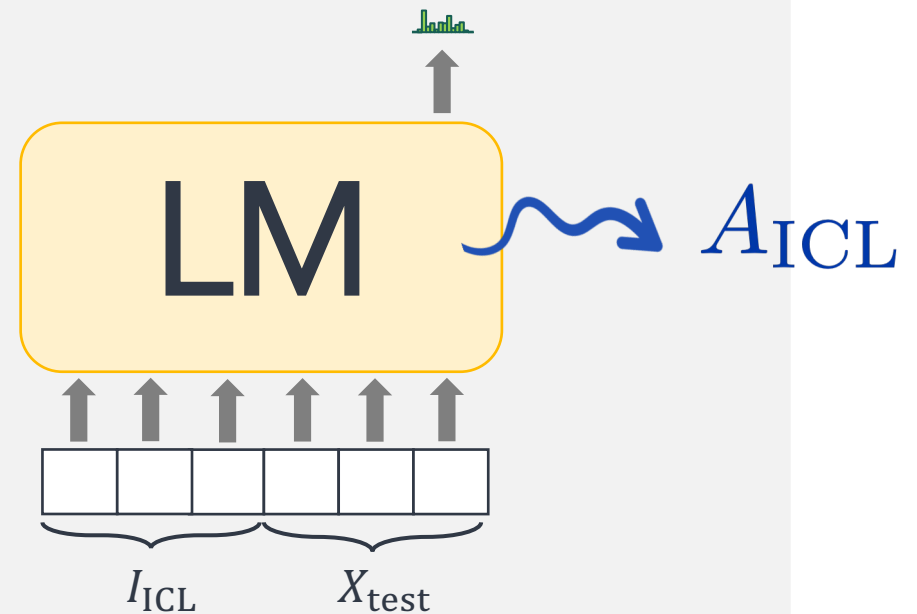


How does ICL compare to SFT?

Supervised Fine-tuning (SFT)



In-Context Learning (ICL)



- Cosine similarity of ICL vs. SFT activations: $\text{sim}(A_{\text{ICL}}, A_{\text{SFT}})$

How does ICL compare to SFT?

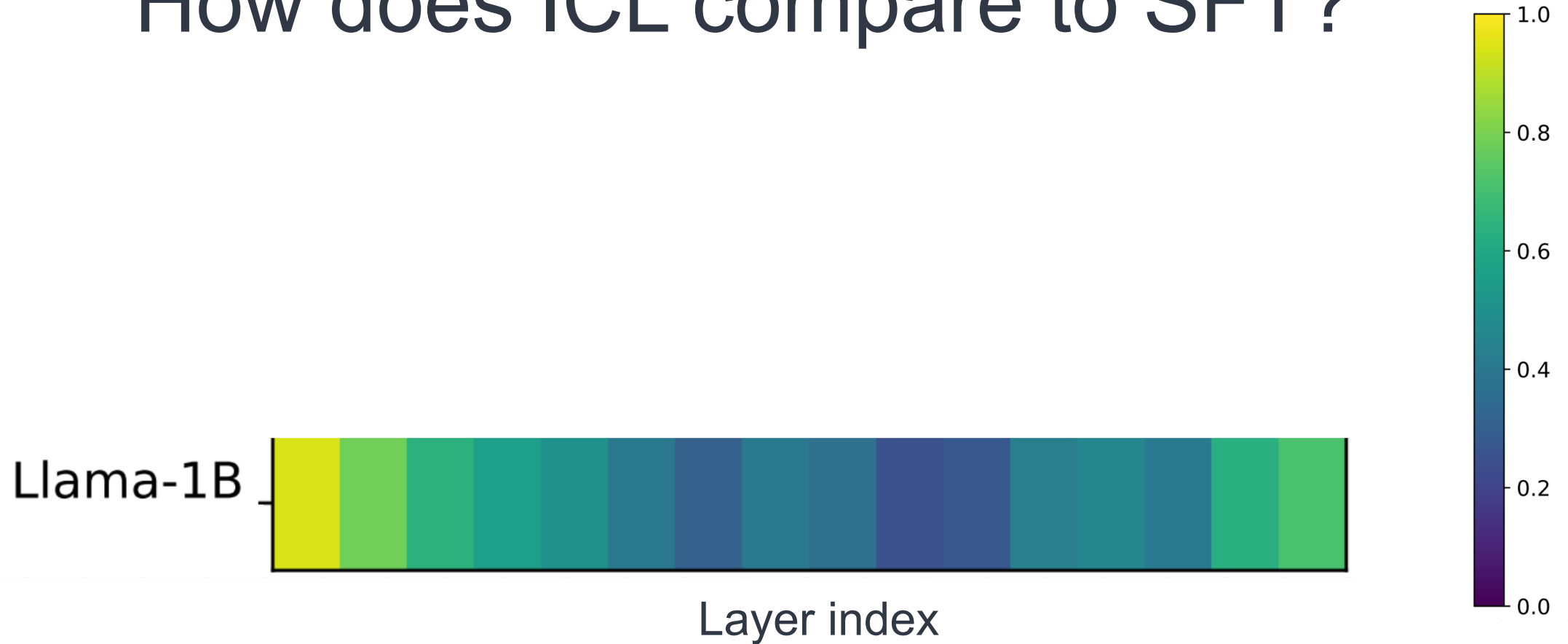
Llama-1B

Layer index



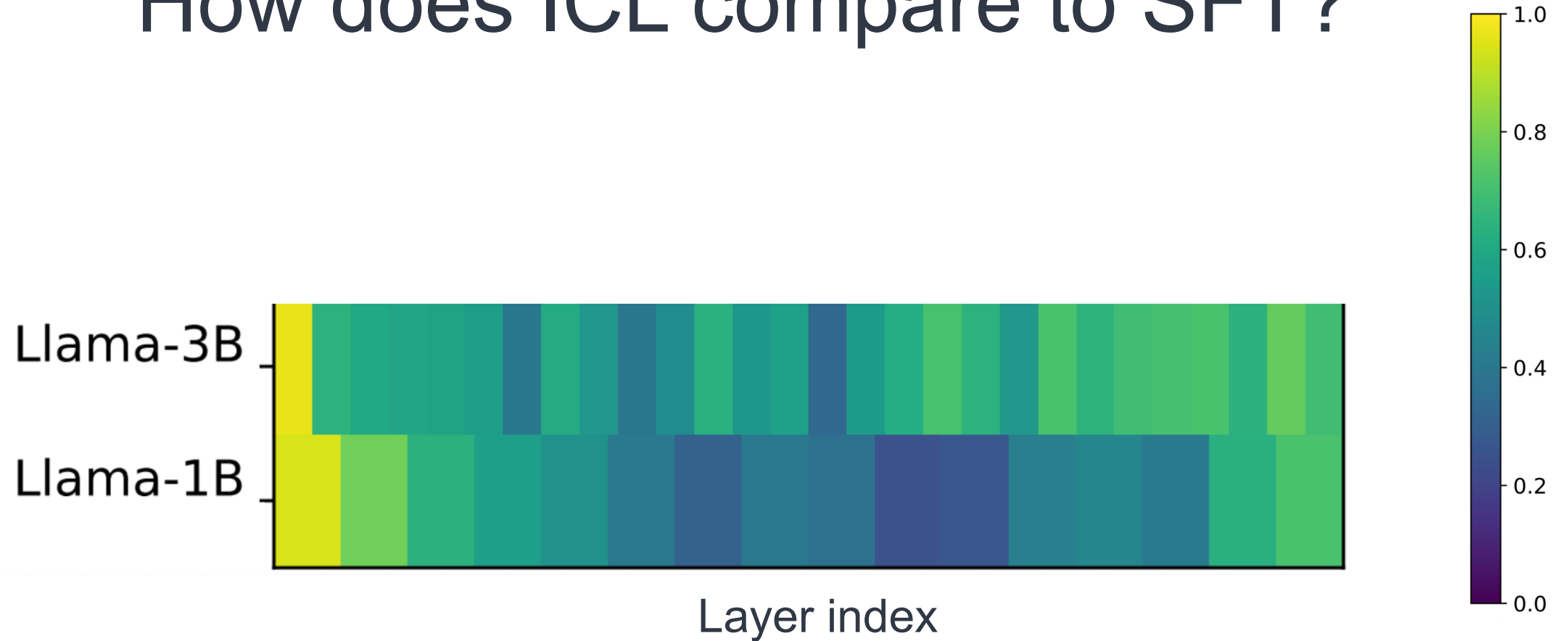
- Cosine similarity of ICL vs. SFT activations: $\text{sim}(A_{\text{ICL}}, A_{\text{SFT}})$

How does ICL compare to SFT?



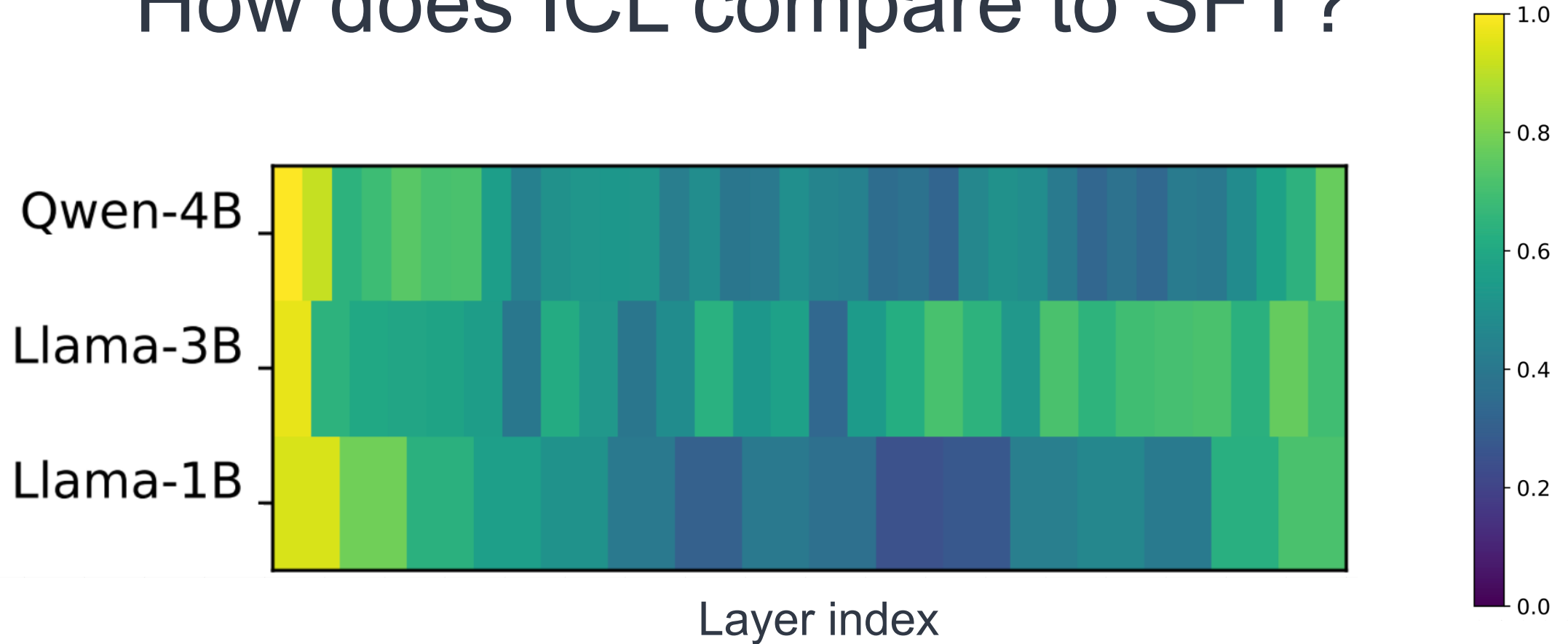
- Cosine similarity of ICL vs. SFT activations: $\text{sim}(A_{\text{ICL}}, A_{\text{SFT}})$

How does ICL compare to SFT?



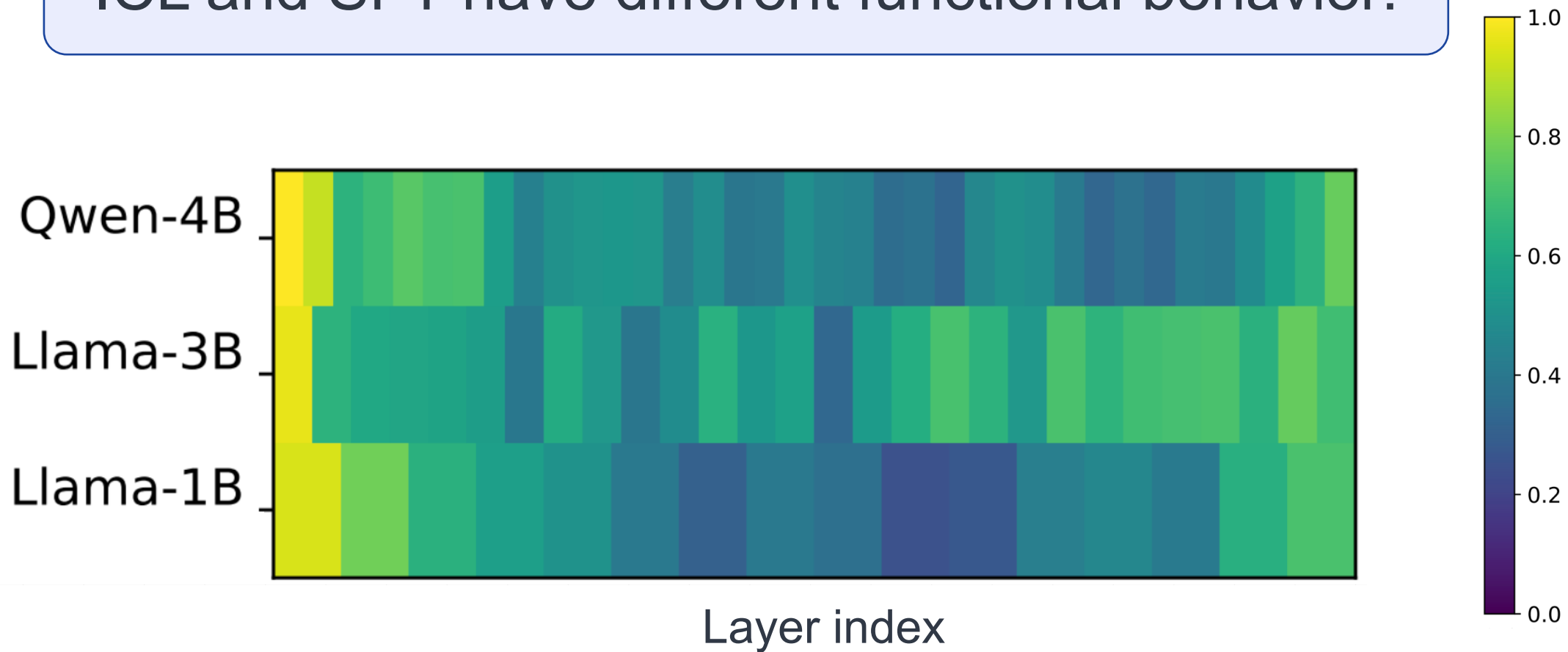
- Cosine similarity of ICL vs. SFT activations: $\text{sim}(A_{\text{ICL}}, A_{\text{SFT}})$

How does ICL compare to SFT?



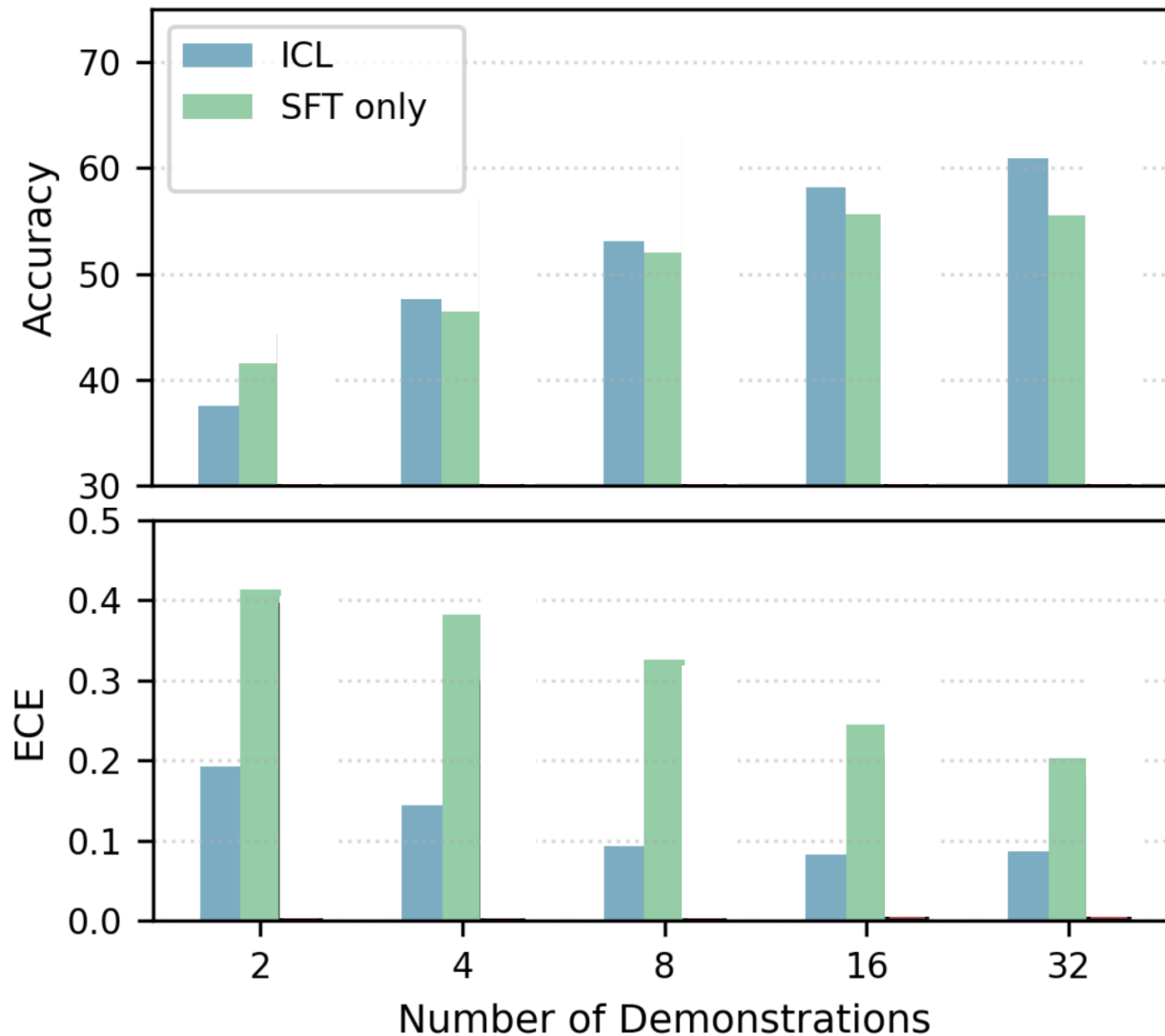
- Cosine similarity of ICL vs. SFT activations: $\text{sim}(A_{\text{ICL}}, A_{\text{SFT}})$

ICL and SFT have different functional behavior!



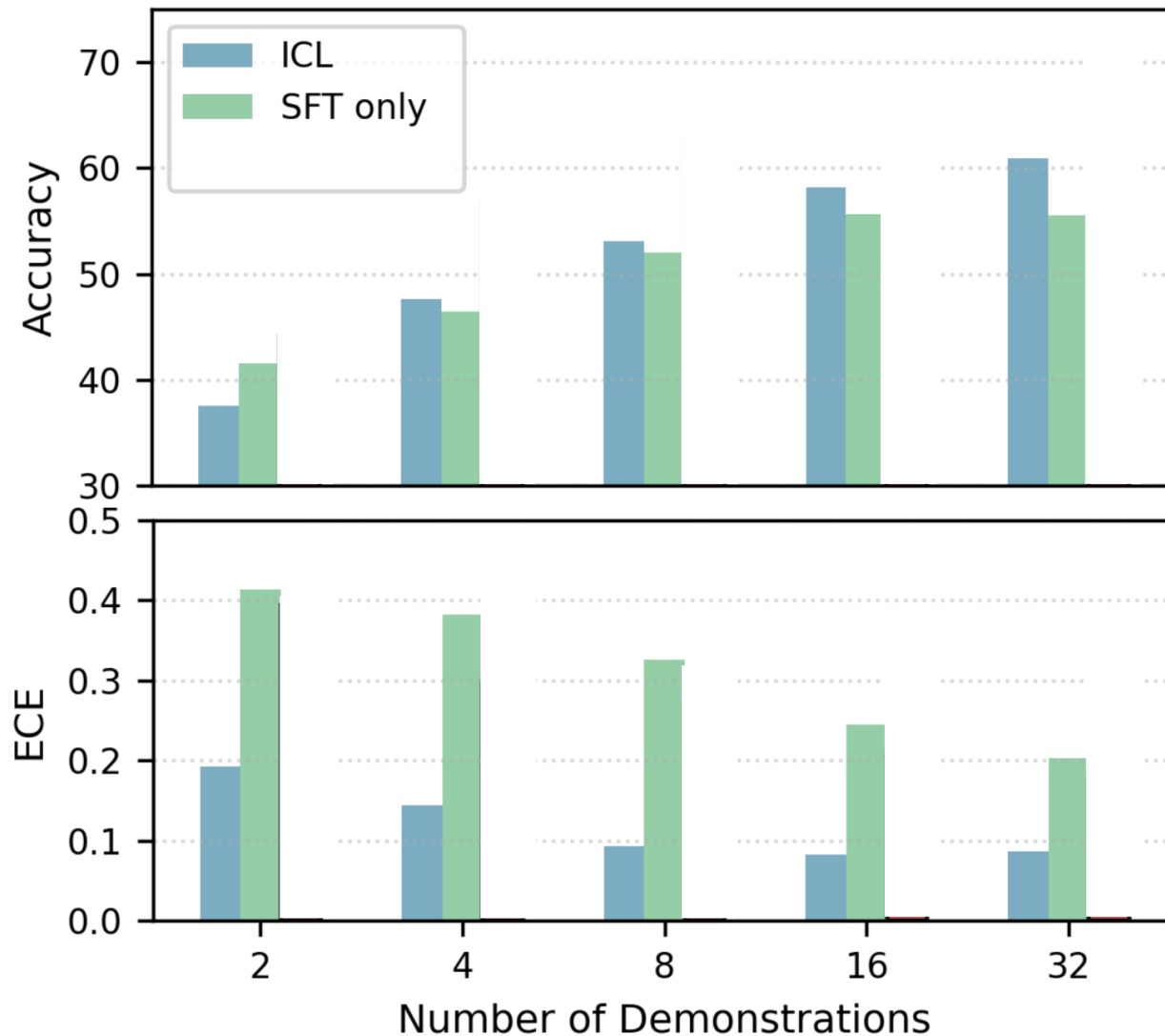
- Cosine similarity of ICL vs. SFT activations: $\text{sim}(A_{\text{ICL}}, A_{\text{SFT}})$

How does ICL compare to SFT?



Other have also made related observations:
(Mosbach et al., 2023; Duan et al., 2024)

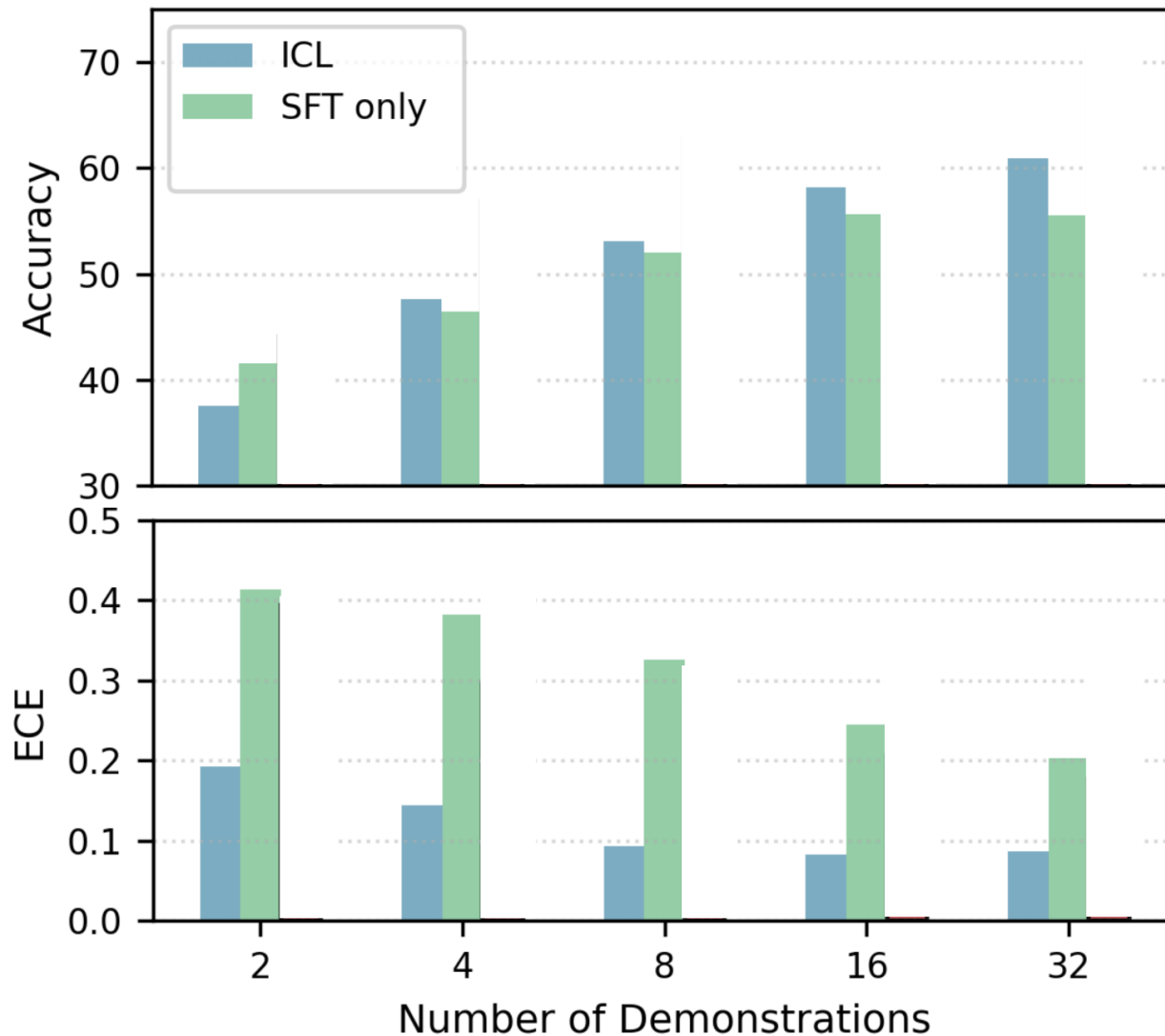
How does ICL compare to SFT?



Similar accuracy.

Other have also made related observations:
(Mosbach et al., 2023; Duan et al., 2024)

How does ICL compare to SFT?

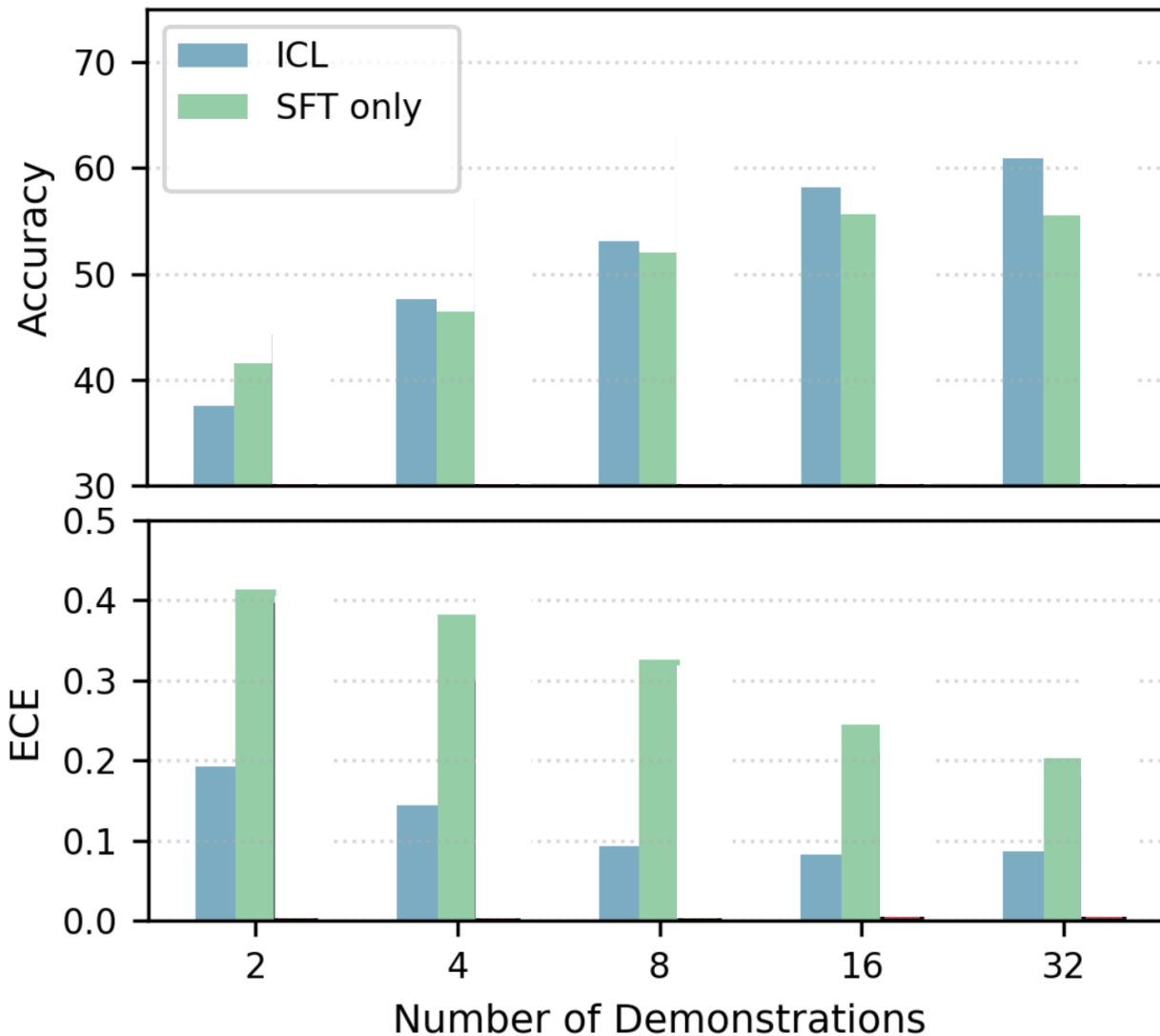


Similar accuracy.

Much better calibration!

Other have also made related observations:
(Mosbach et al., 2023; Duan et al., 2024)

SPOILER ALERT

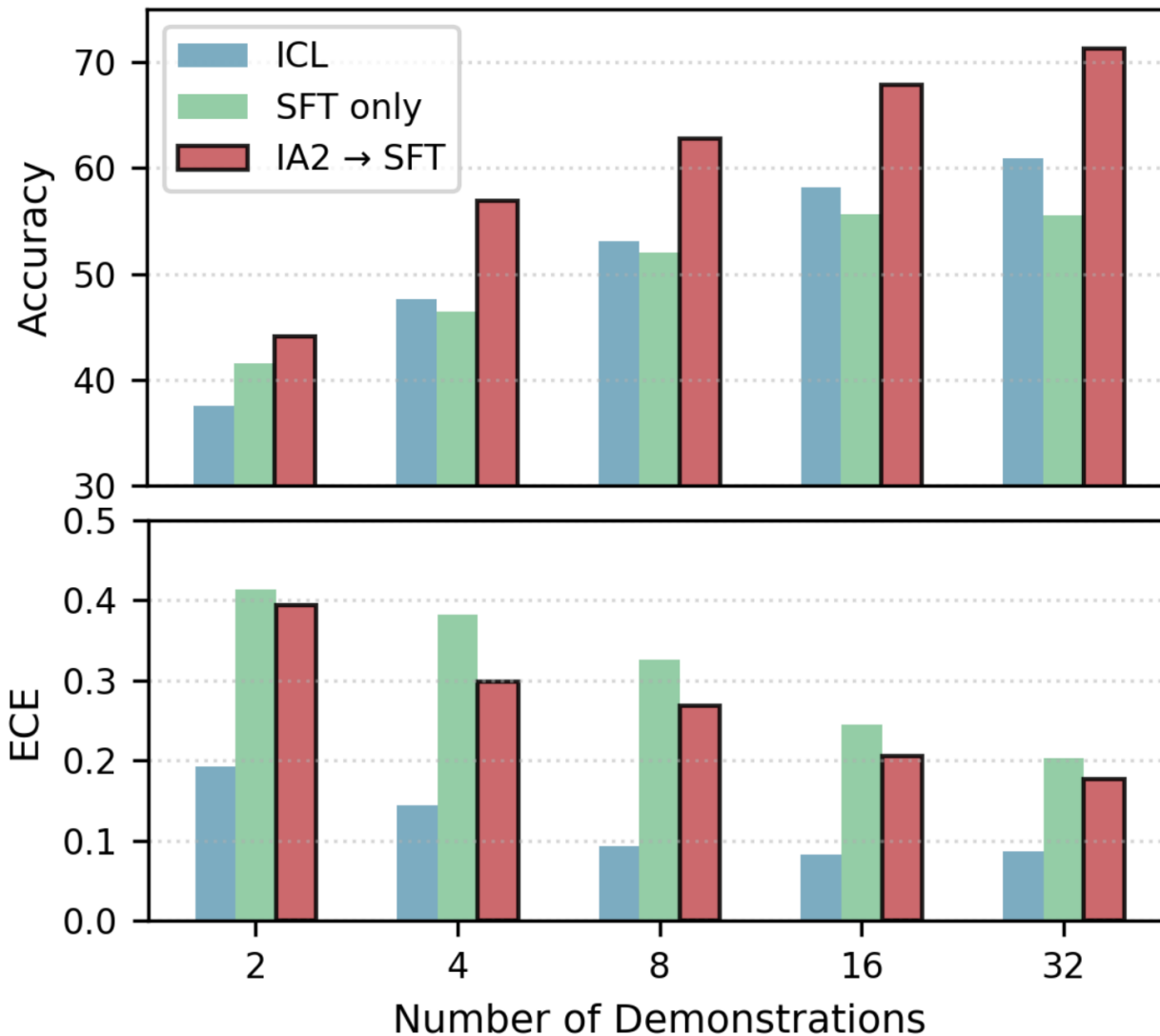


Similar accuracy.

Much better calibration!

Other have also made related observations:
(Mosbach et al., 2023; Duan et al., 2024)

SPOILER ALERT



Similar accuracy.

Much better calibration!

Other have also made related observations:
(Mosbach et al., 2023; Duan et al., 2024)

Alignment with ICL Activations (IA2)

- **Main idea:** Adjusting model **weights** so its **activations** are aligned with that of an **ICL-conditioned** model.



Alignment with ICL Activations (IA2)

- **Step 1:** Collect information-rich ICL activations

Alignment with ICL Activations (IA2)

- **Step 1:** Collect information-rich ICL activations

Given some labeled data

task $\mathcal{T} : \mathcal{X} \rightarrow \mathcal{Y}$

$$\left\{ \begin{array}{l} X^1 \rightarrow Y^1 \\ \vdots \\ X^N \rightarrow Y^N \end{array} \right.$$

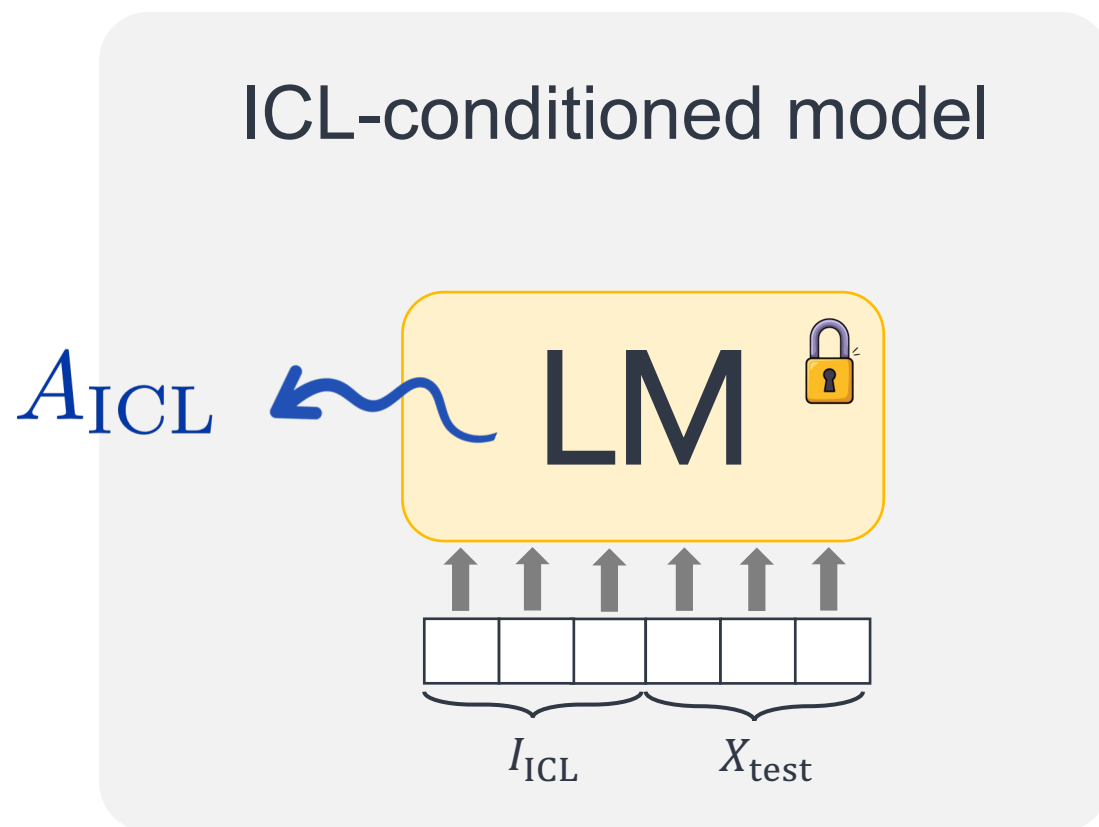
Alignment with ICL Activations (IA2)

- **Step 1:** Collect information-rich ICL activations

Given some labeled data

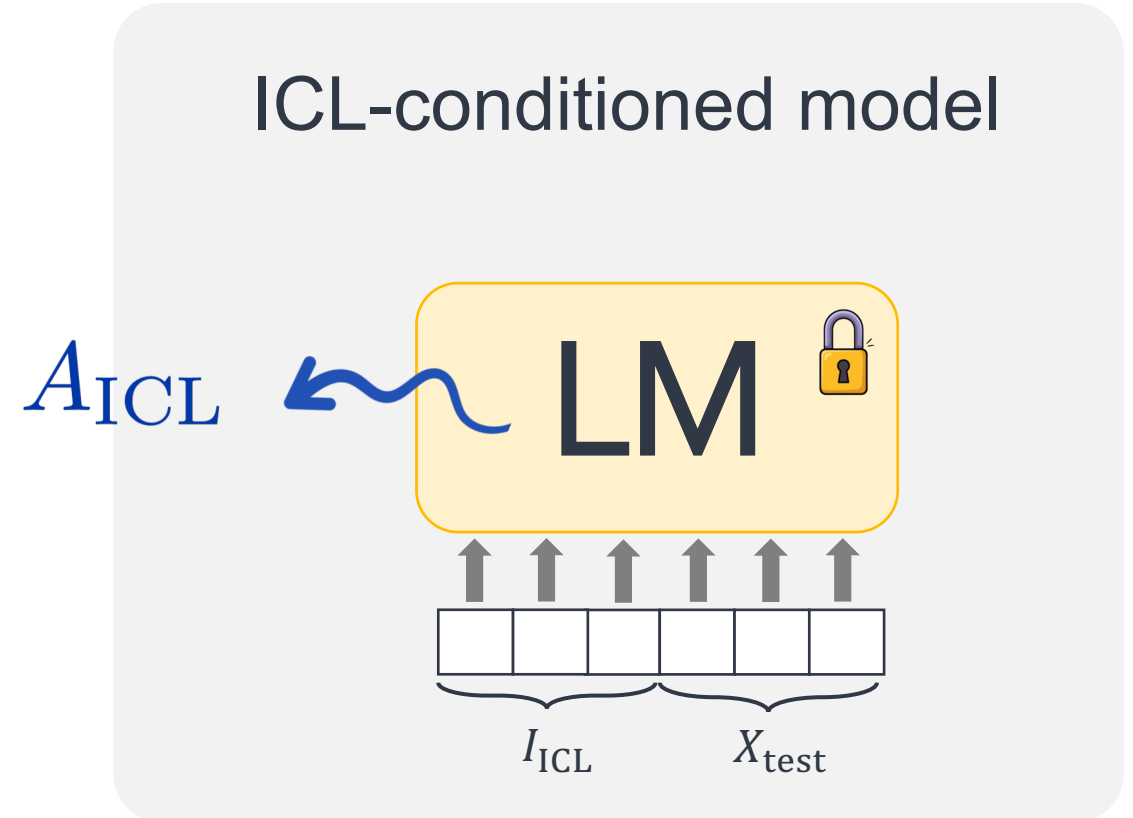
task $\mathcal{T} : \mathcal{X} \rightarrow \mathcal{Y}$

$$\left\{ \begin{array}{l} X^1 \rightarrow Y^1 \\ \vdots \\ X^N \rightarrow Y^N \end{array} \right.$$



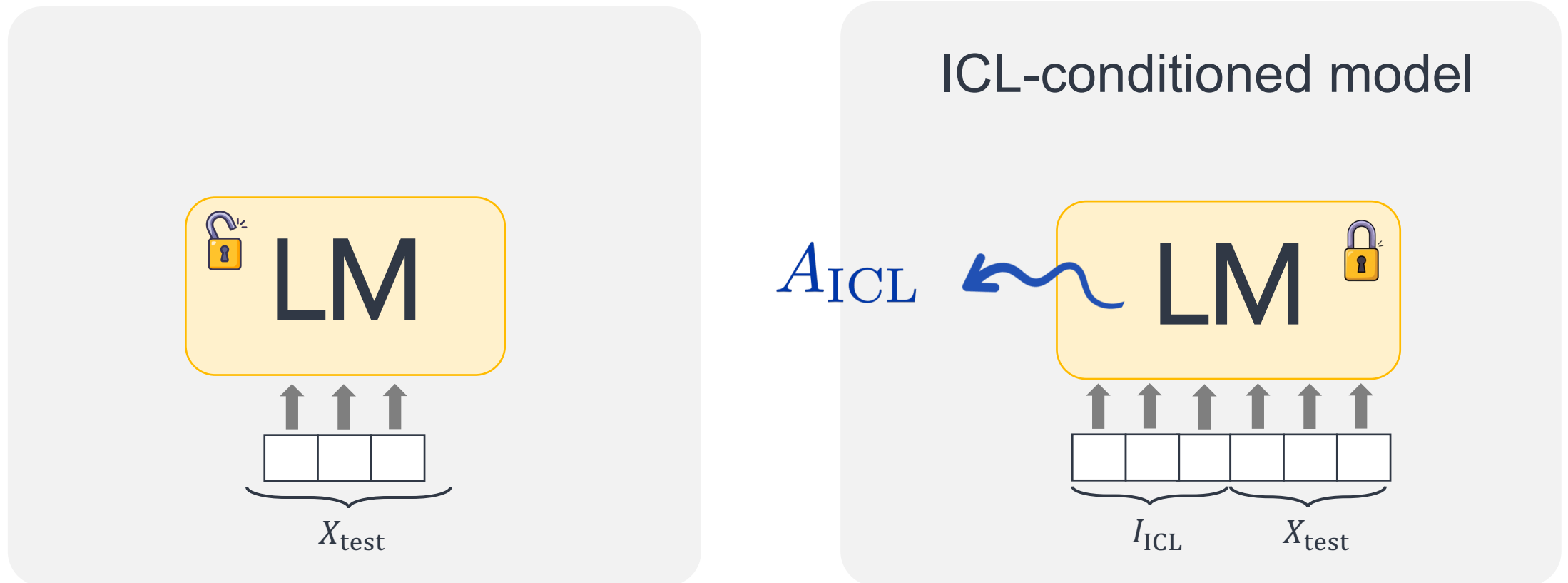
Alignment with ICL Activations (IA2)

- **Step 2:** Perform activation alignment



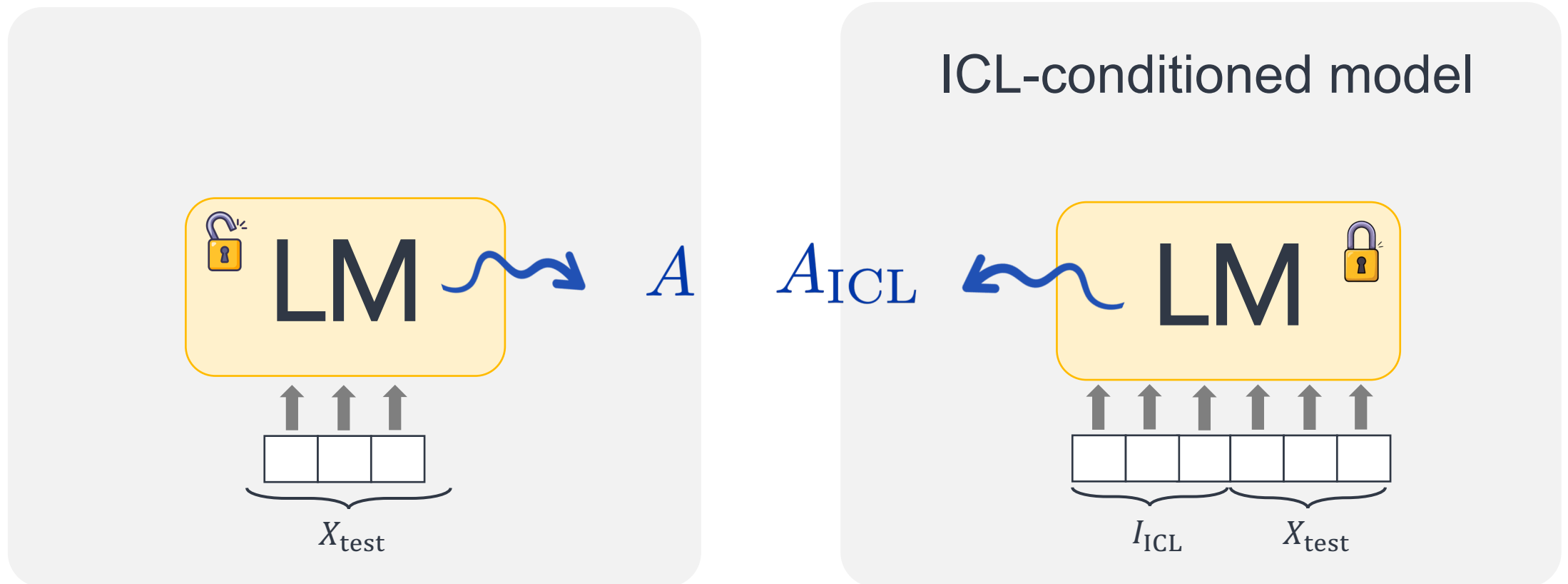
Alignment with ICL Activations (IA2)

- **Step 2:** Perform activation alignment



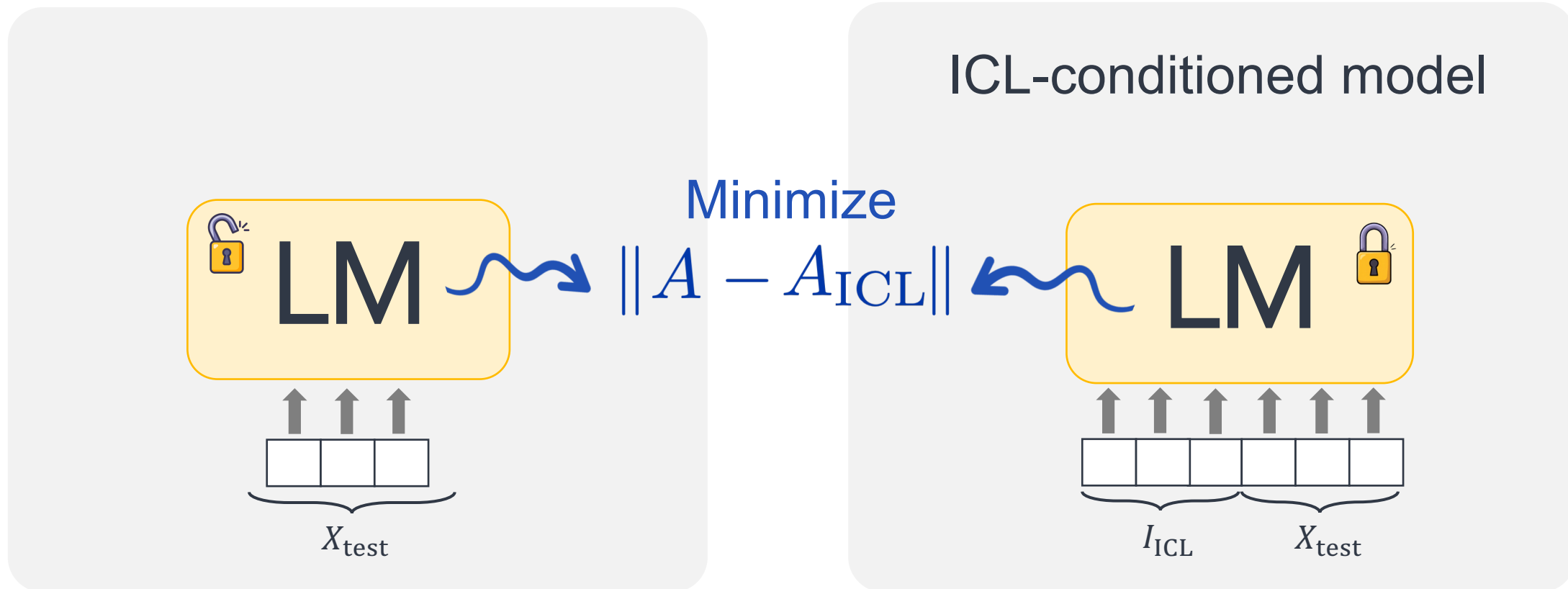
Alignment with ICL Activations (IA2)

- **Step 2:** Perform activation alignment



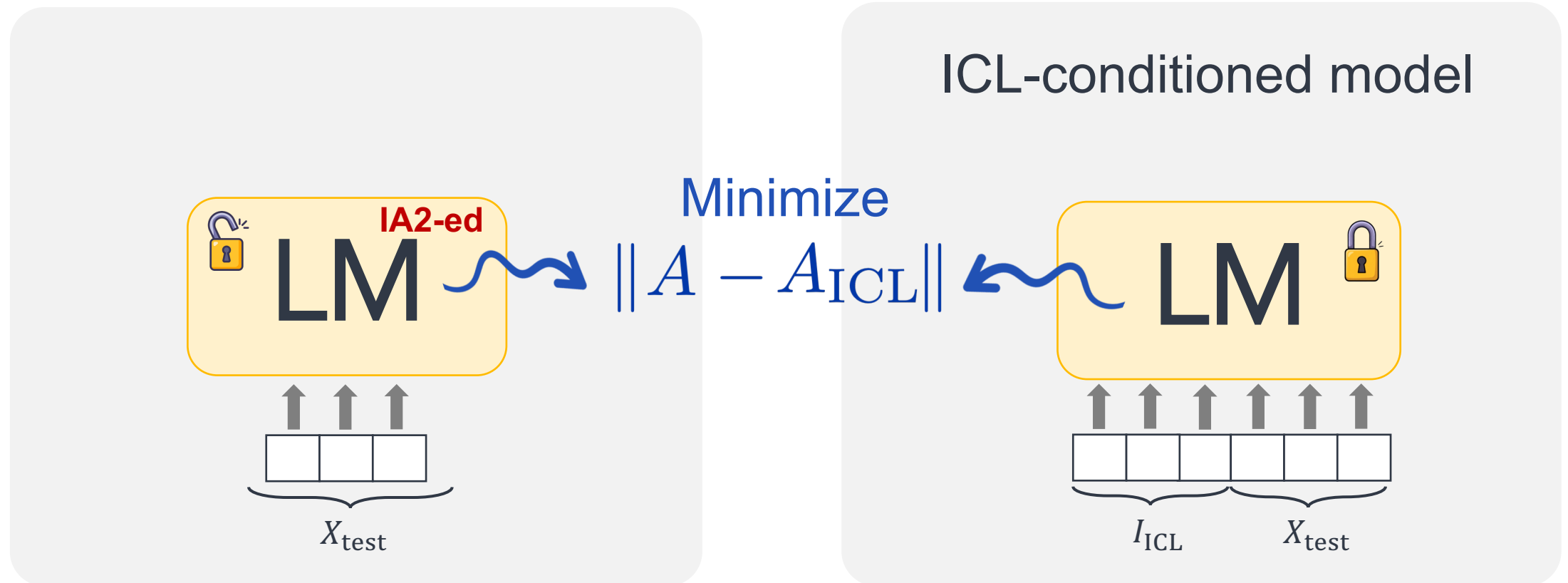
Alignment with ICL Activations (IA2)

- **Step 2:** Perform activation alignment



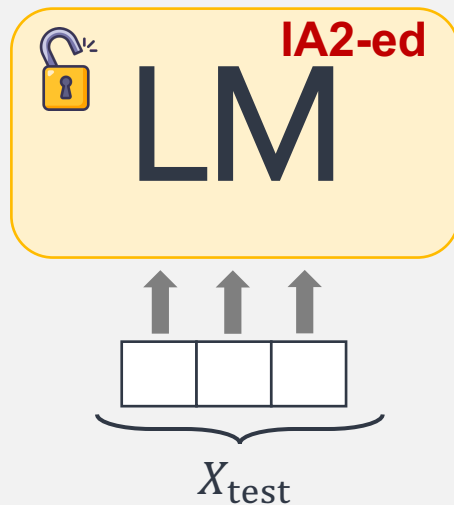
Alignment with ICL Activations (IA2)

- **Step 2:** Perform activation alignment



Alignment with ICL Activations (IA2)

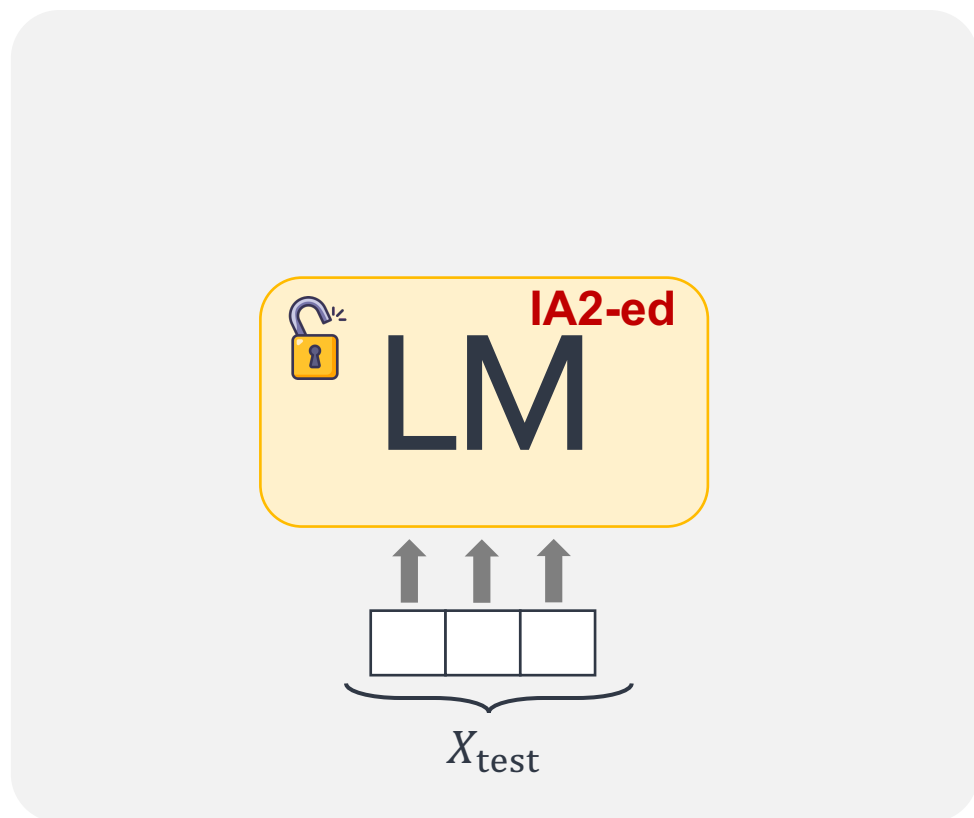
- **Step 3: Perform SFT**



Alignment with ICL Activations (IA2)

- **Step 3: Perform SFT**

Using **the same** labeled data

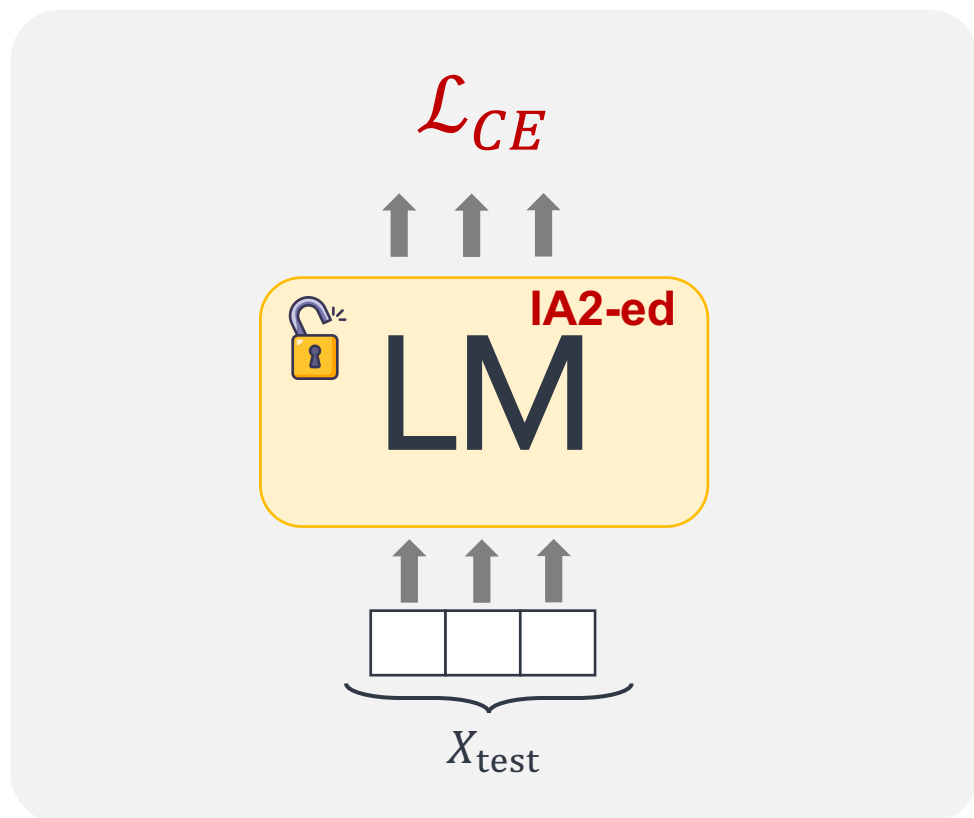


$$\left\{ \begin{array}{l} X^1 \rightarrow Y^1 \\ \vdots \\ X^N \rightarrow Y^N \end{array} \right.$$

Alignment with ICL Activations (IA2)

- **Step 3: Perform SFT**

Using **the same** labeled data



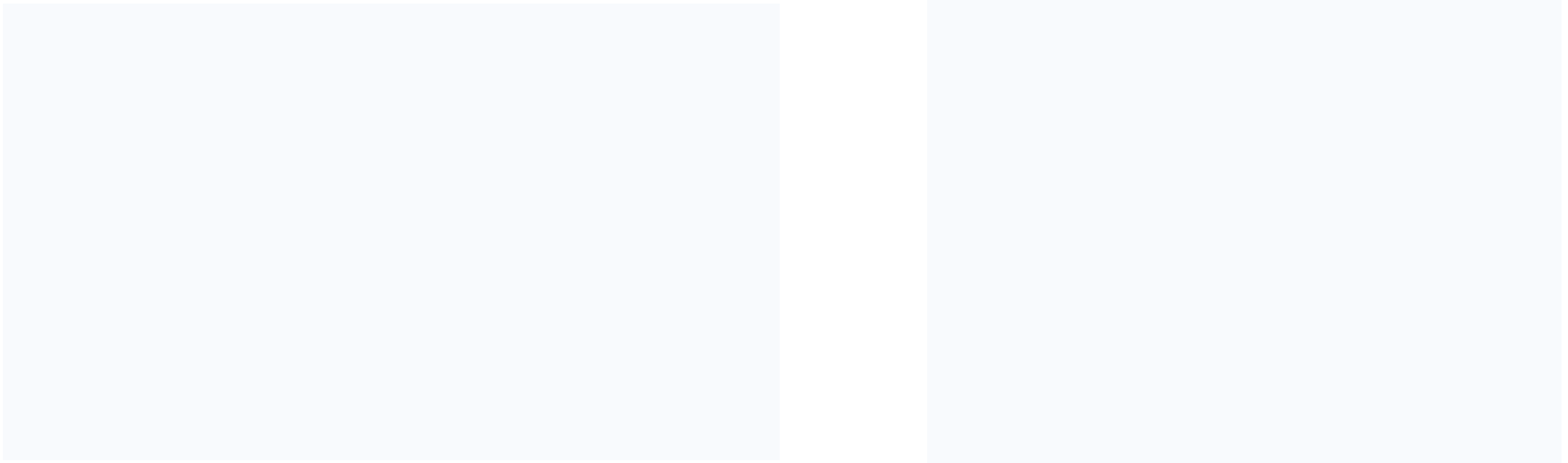
$$\left\{ \begin{array}{l} X^1 \rightarrow Y^1 \\ \vdots \\ X^N \rightarrow Y^N \end{array} \right.$$

Minimize \mathcal{L}_{CE}

Experimental Results: Accuracy ↑

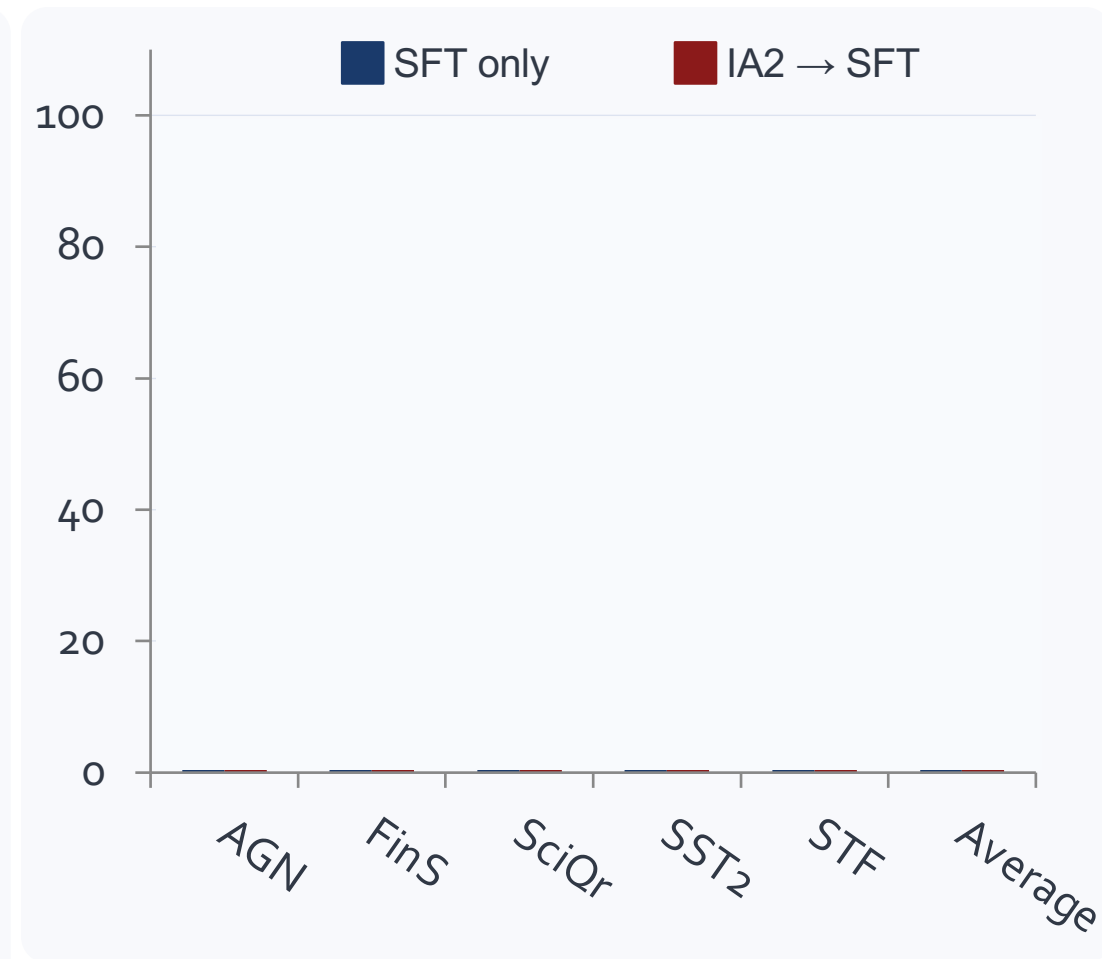
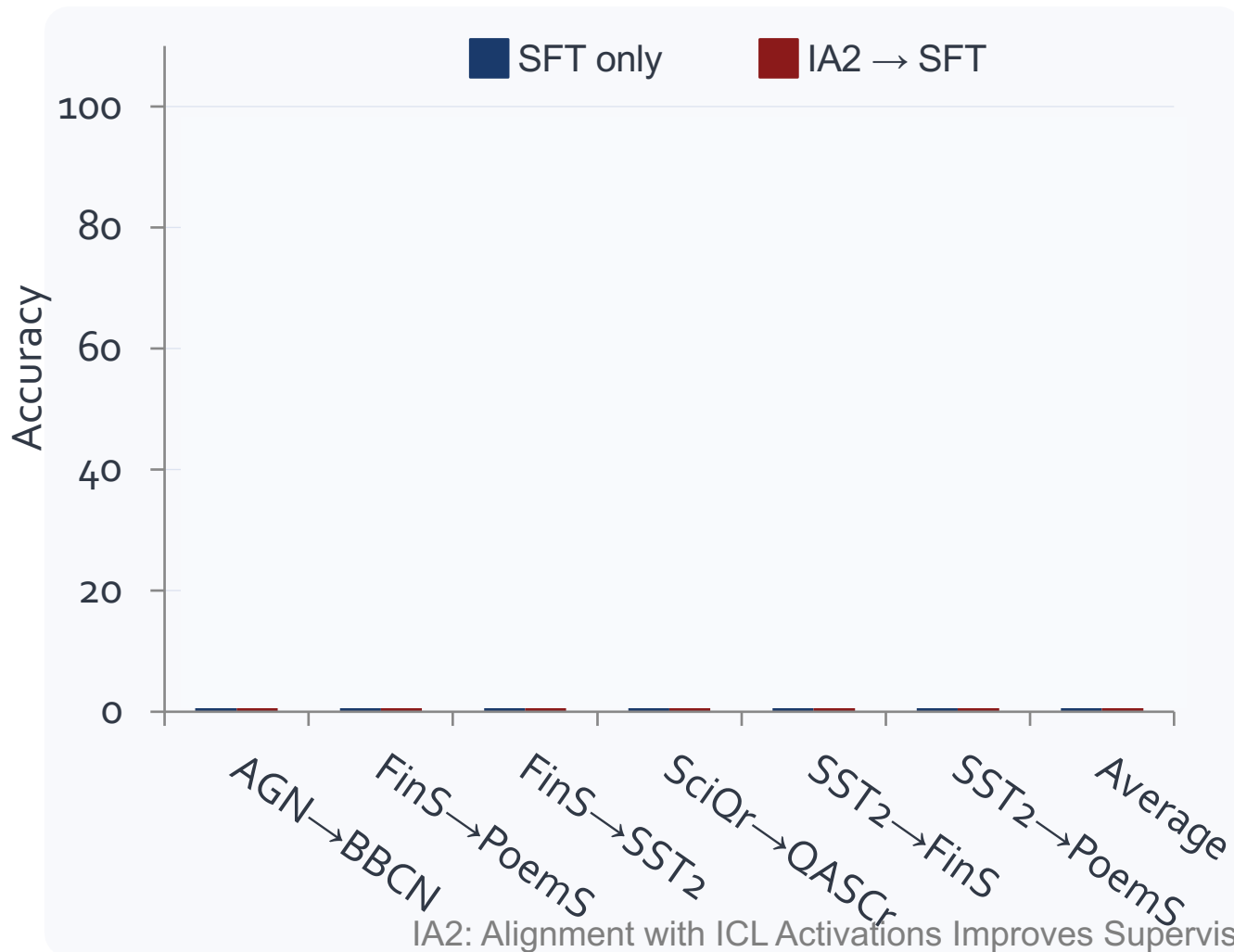
- Model: Qwen3-4B-Base, Number of labeled data: $N = 4$.

Accuracy



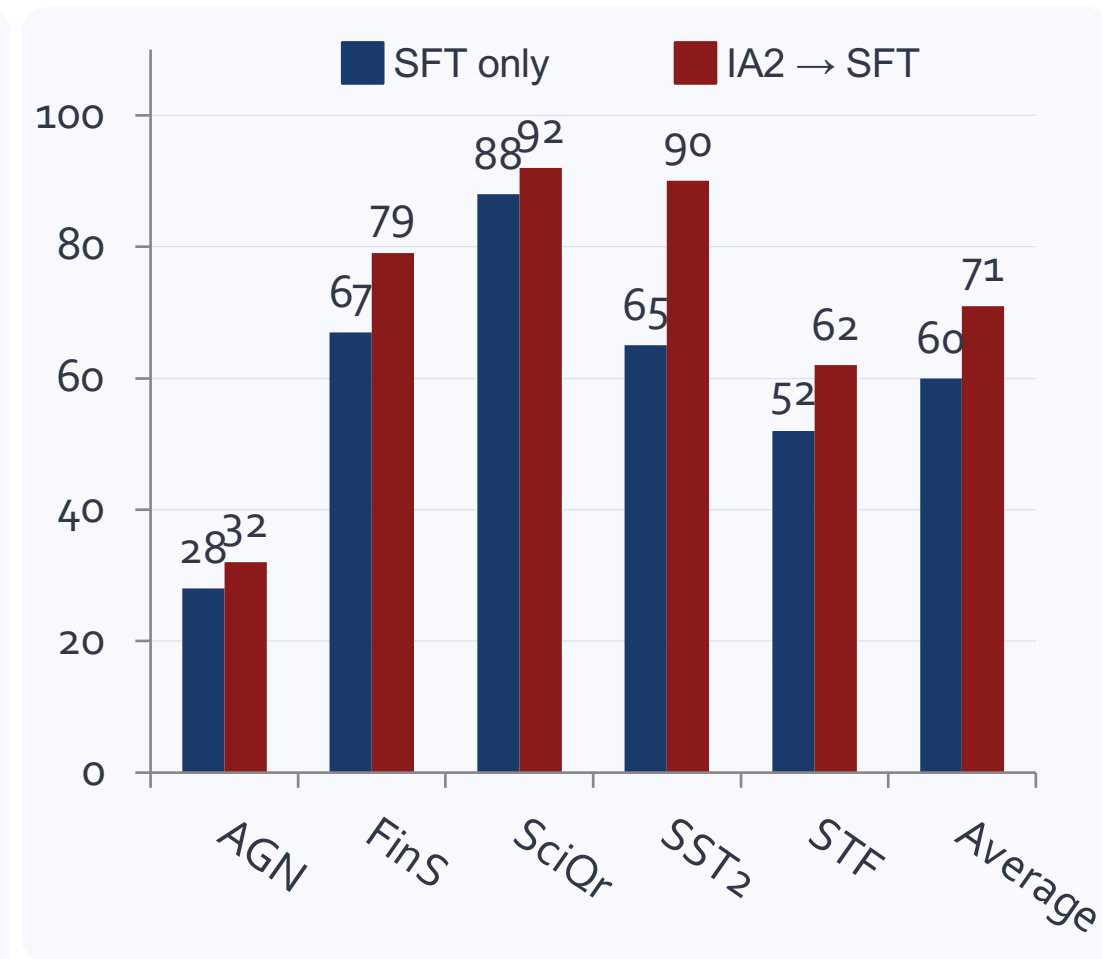
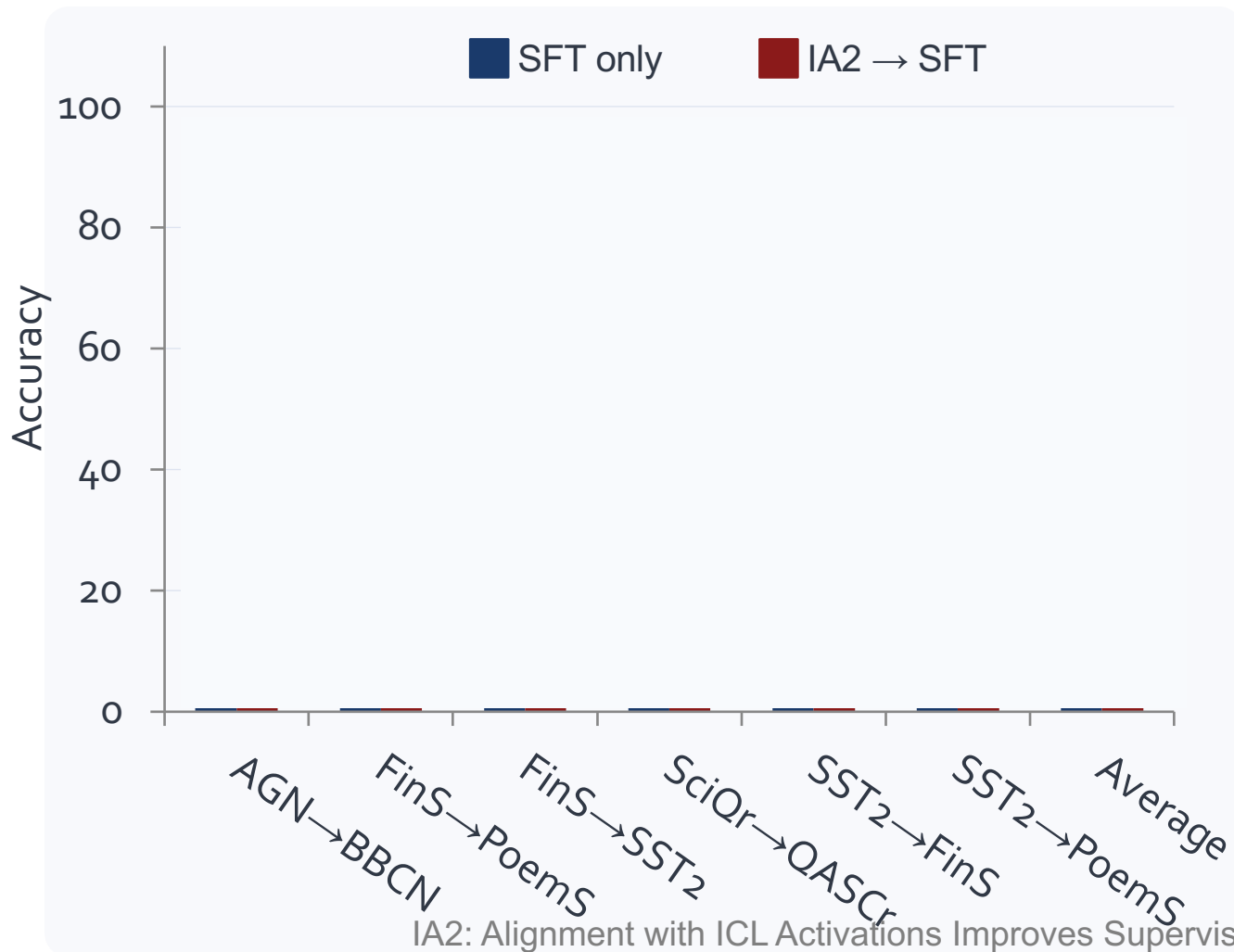
Experimental Results: Accuracy ↑

- Model: Qwen3-4B-Base, Number of labeled data: $N = 4$.



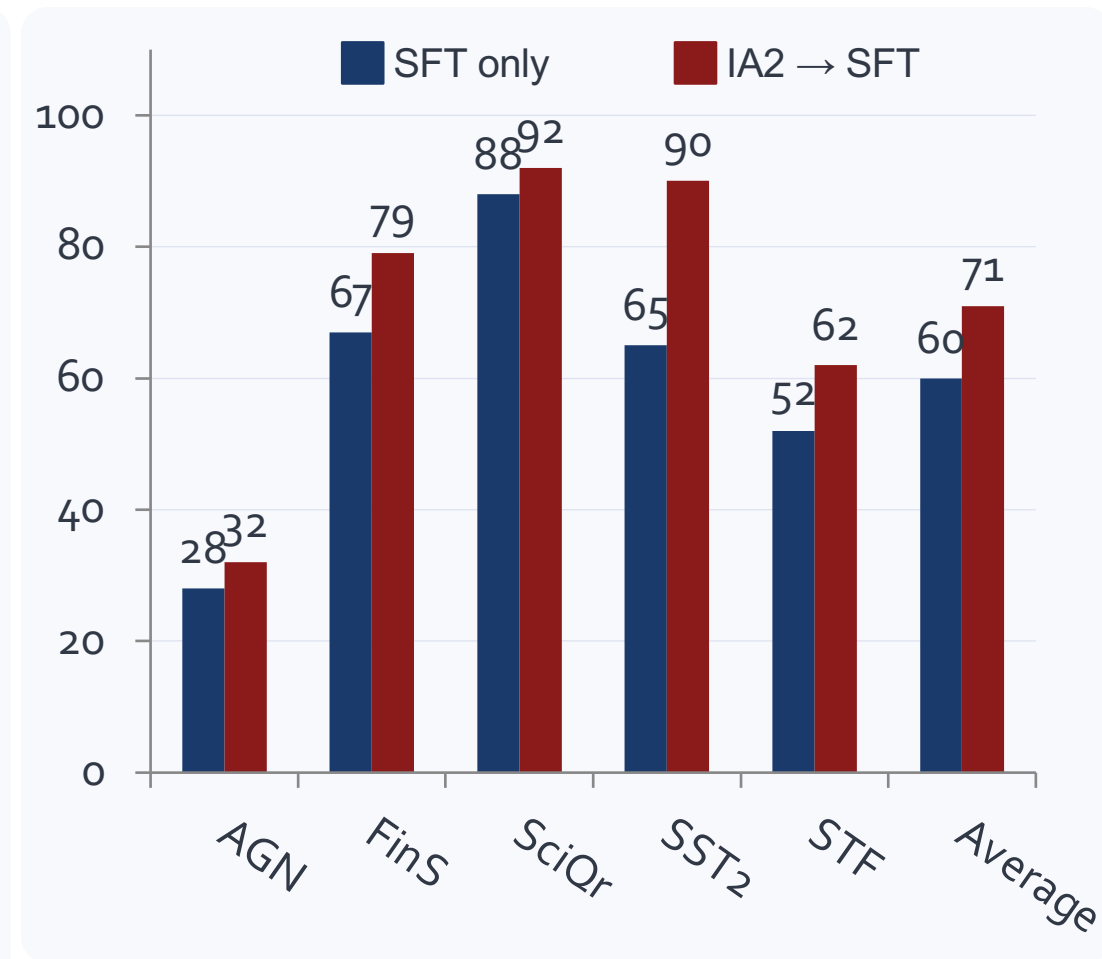
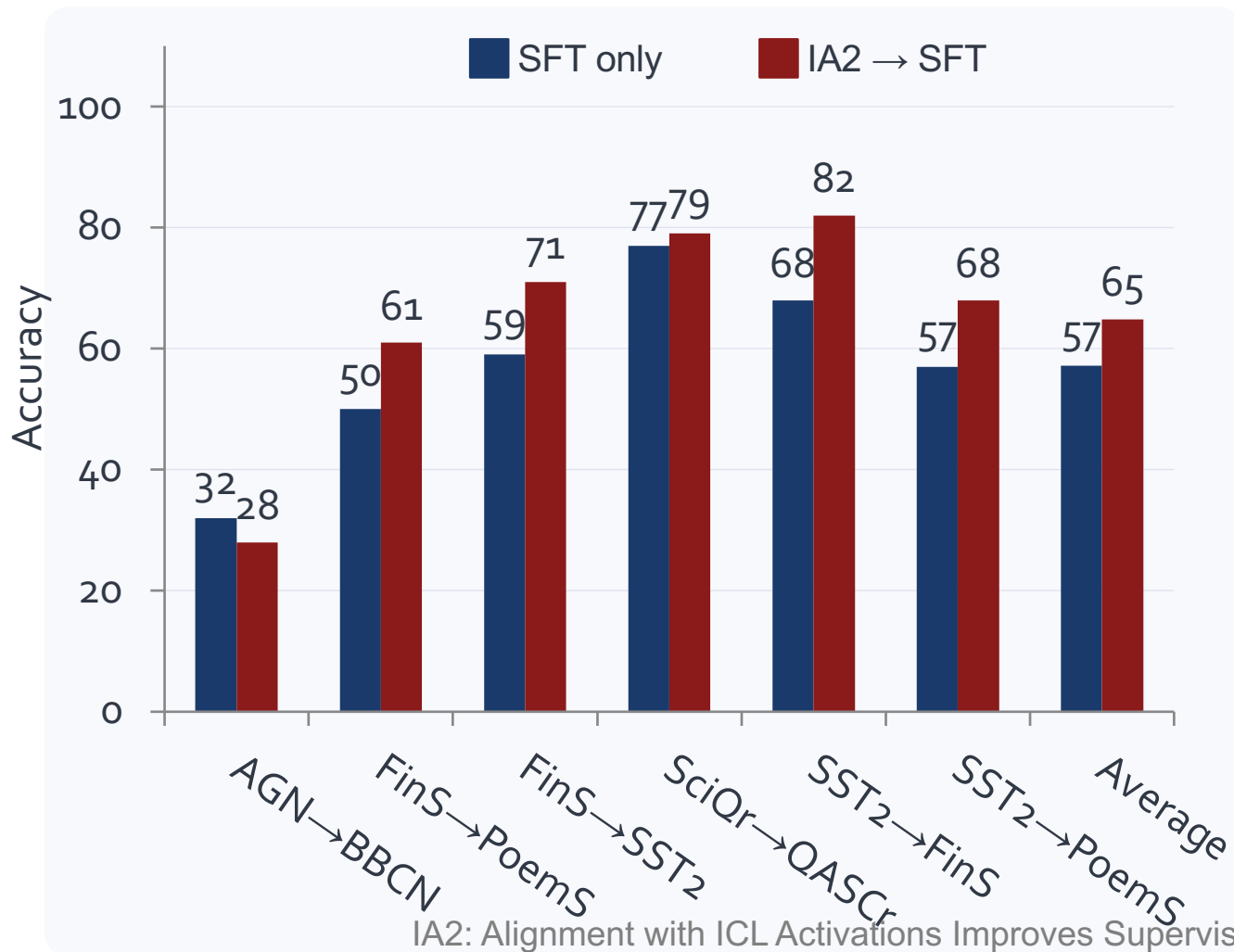
Experimental Results: Accuracy ↑

- Model: Qwen3-4B-Base, Number of labeled data: $N = 4$.



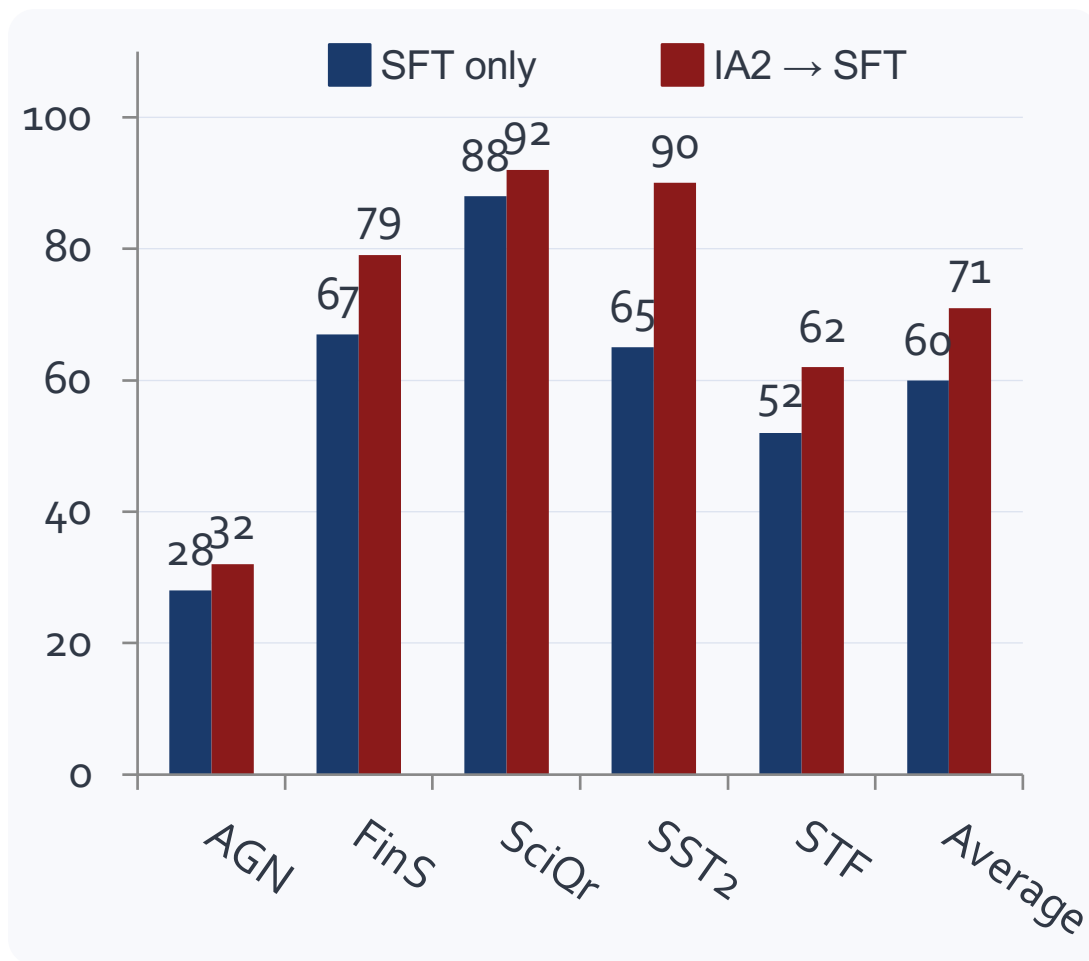
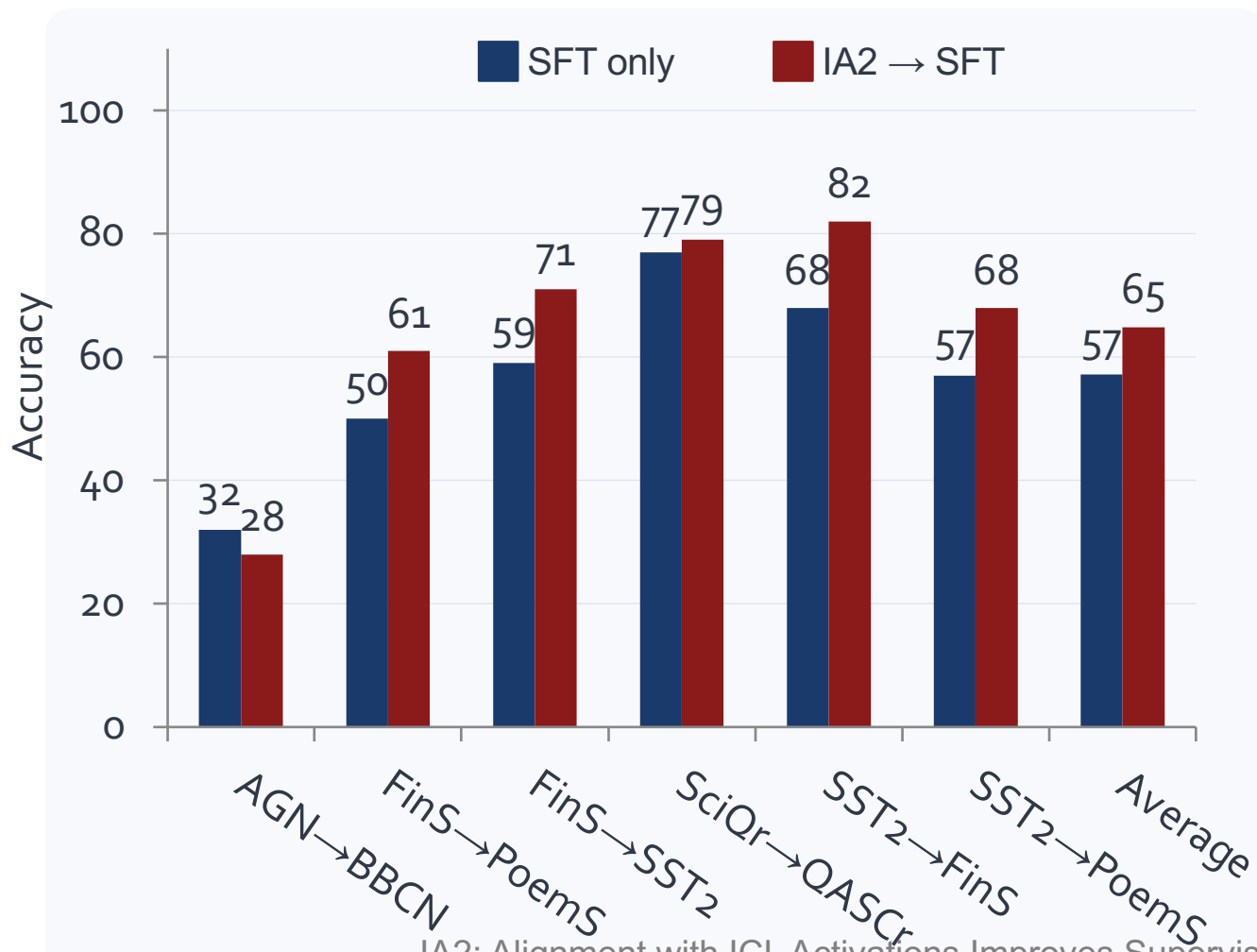
Experimental Results: Accuracy ↑

- Model: Qwen3-4B-Base, Number of labeled data: $N = 4$.



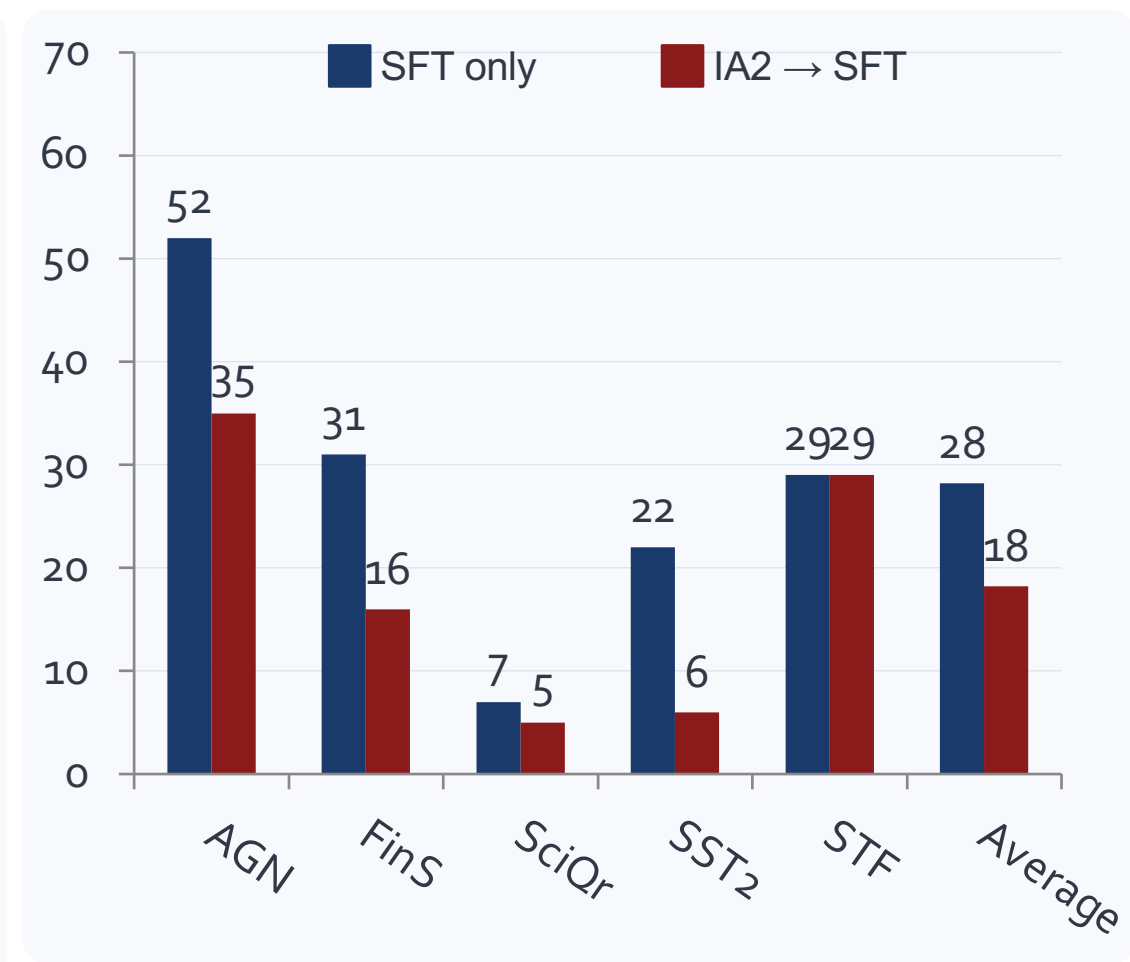
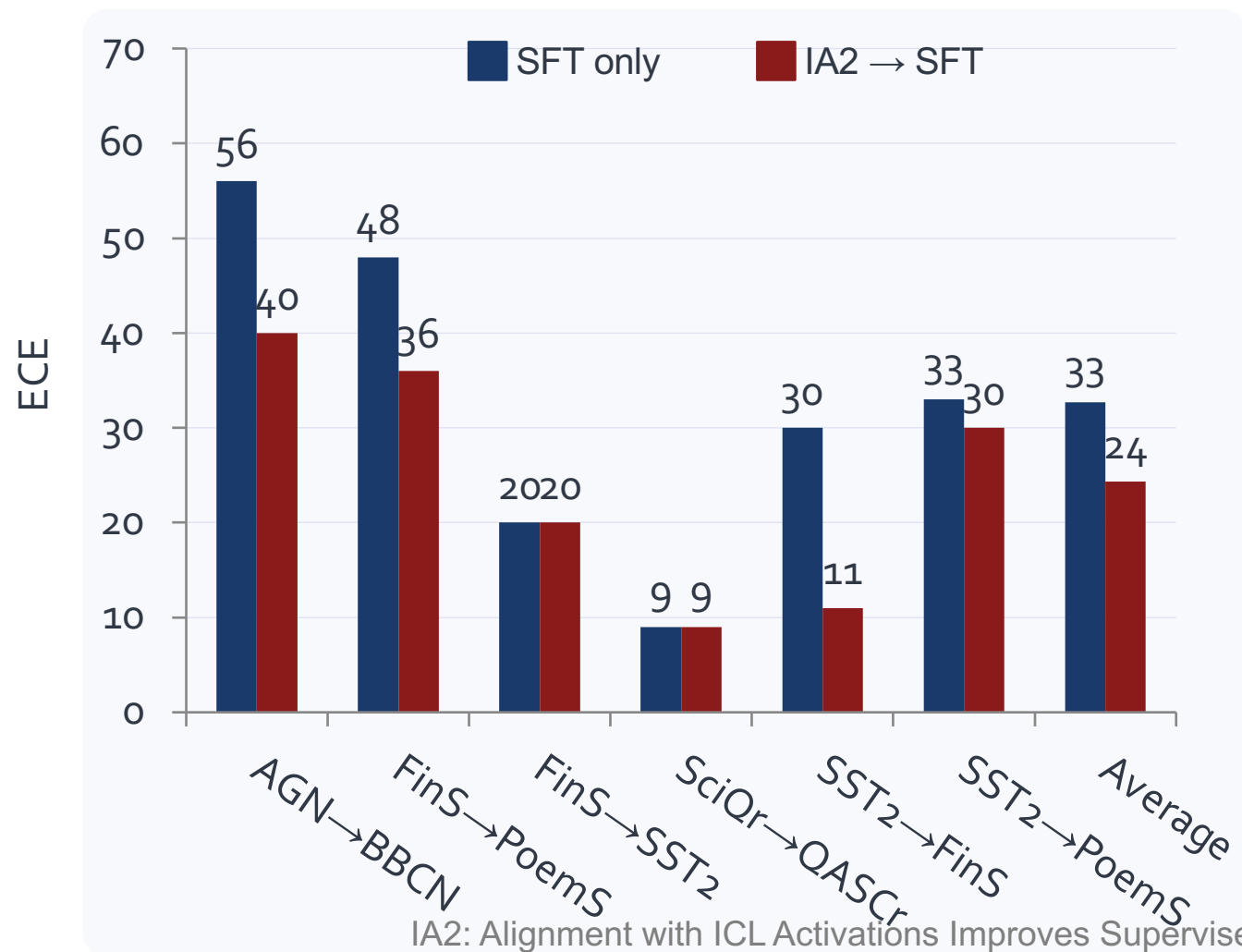
Pre-conditioning w/ IA2 improves accuracy!

- Model: Qwen3-4B-Base, Number of labeled data: $N = 4$.



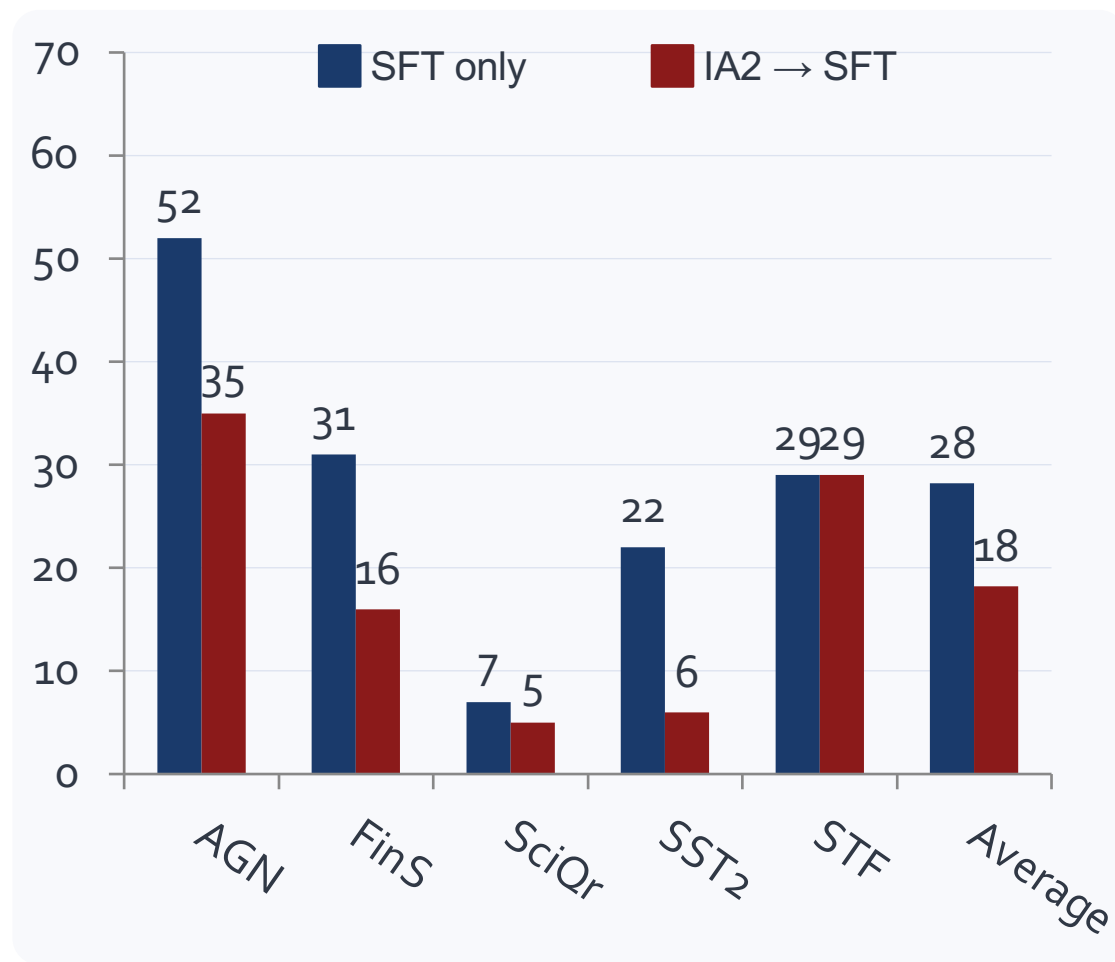
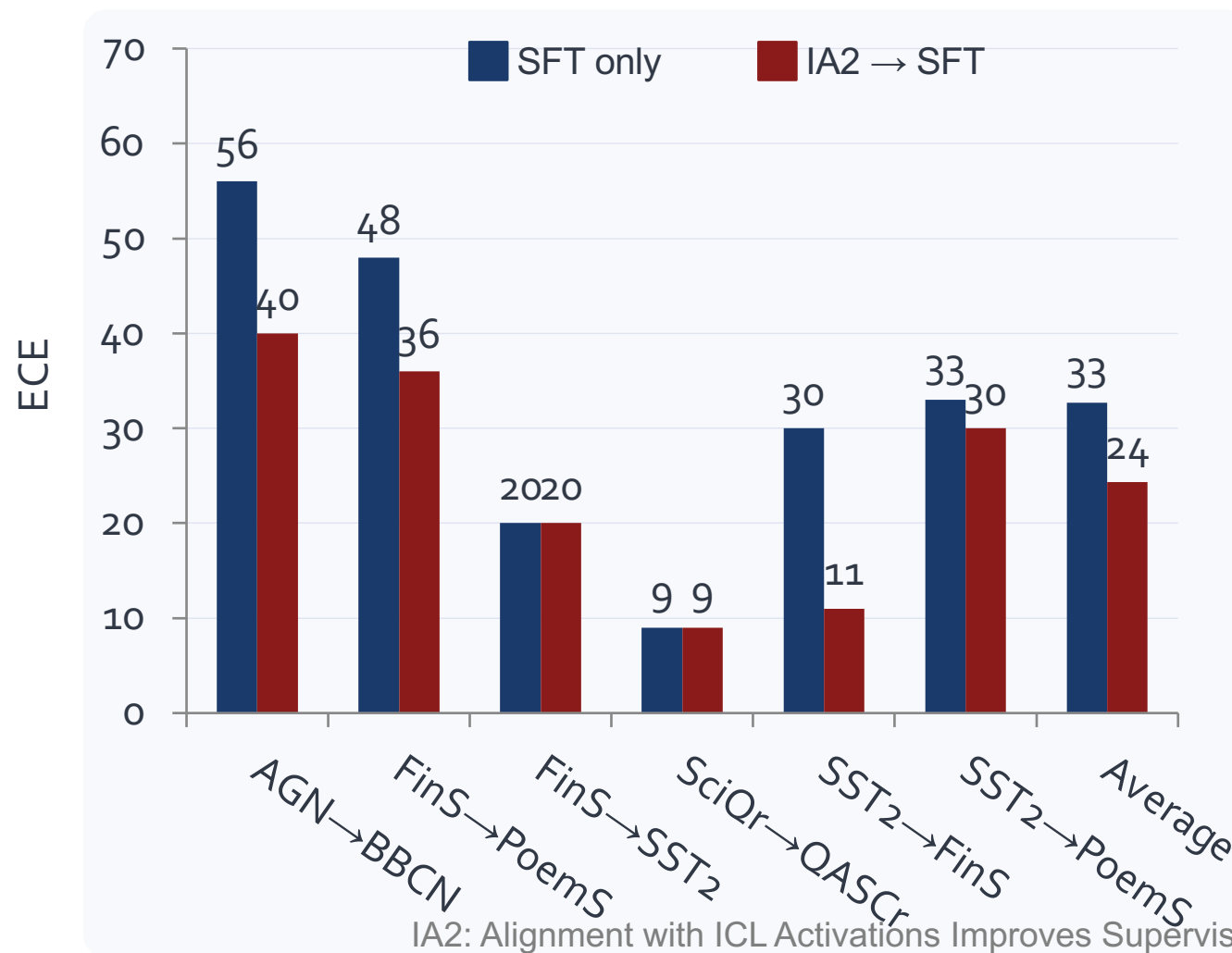
Experimental Results: Expected Calibration Error ↓

- Model: Qwen3-4B-Base, Number of labeled data: $N = 4$.



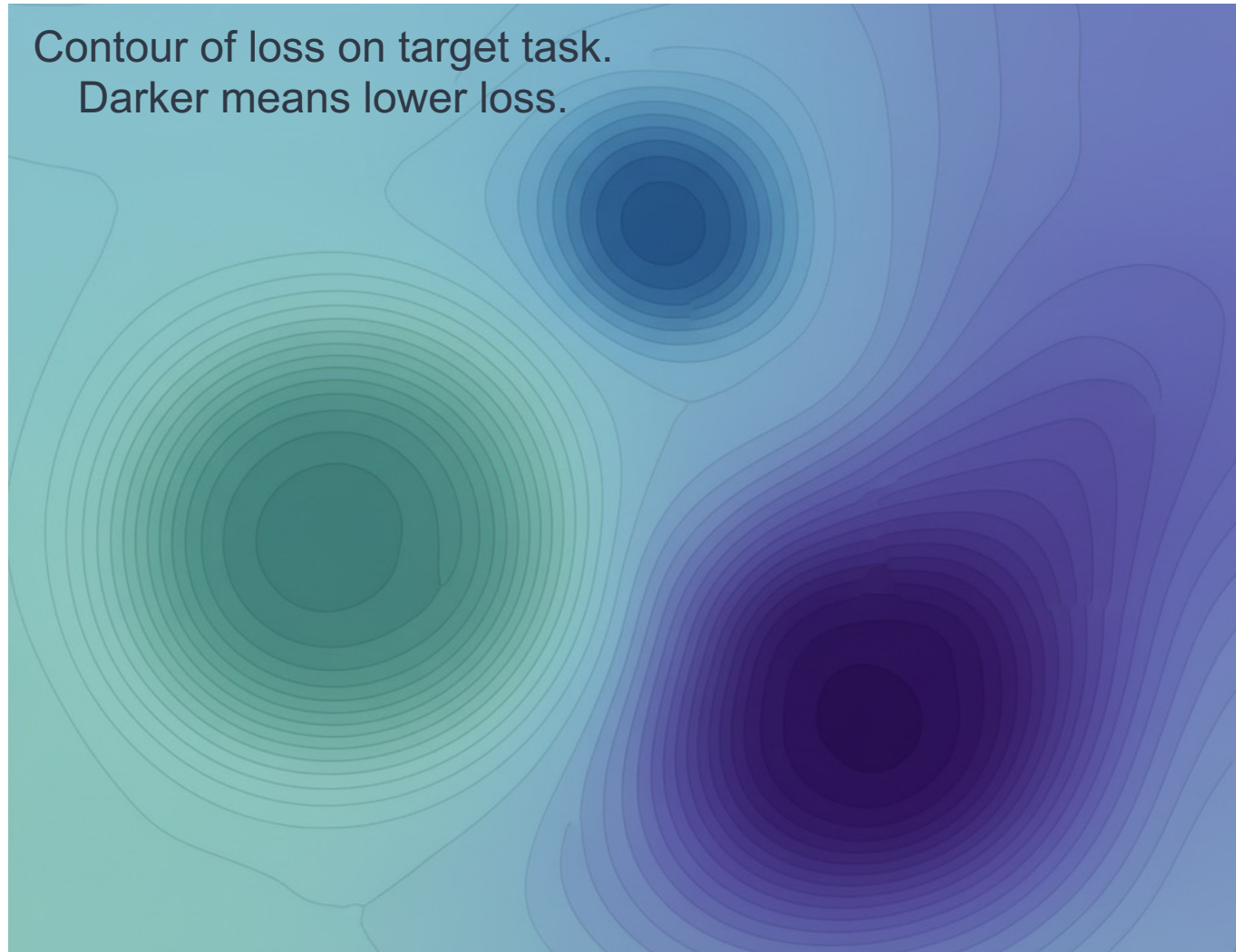
Pre-conditioning w/ IA2 improves accuracy and calibration!

- Model: Qwen3-4B-Base, Number of labeled data: $N = 4$.

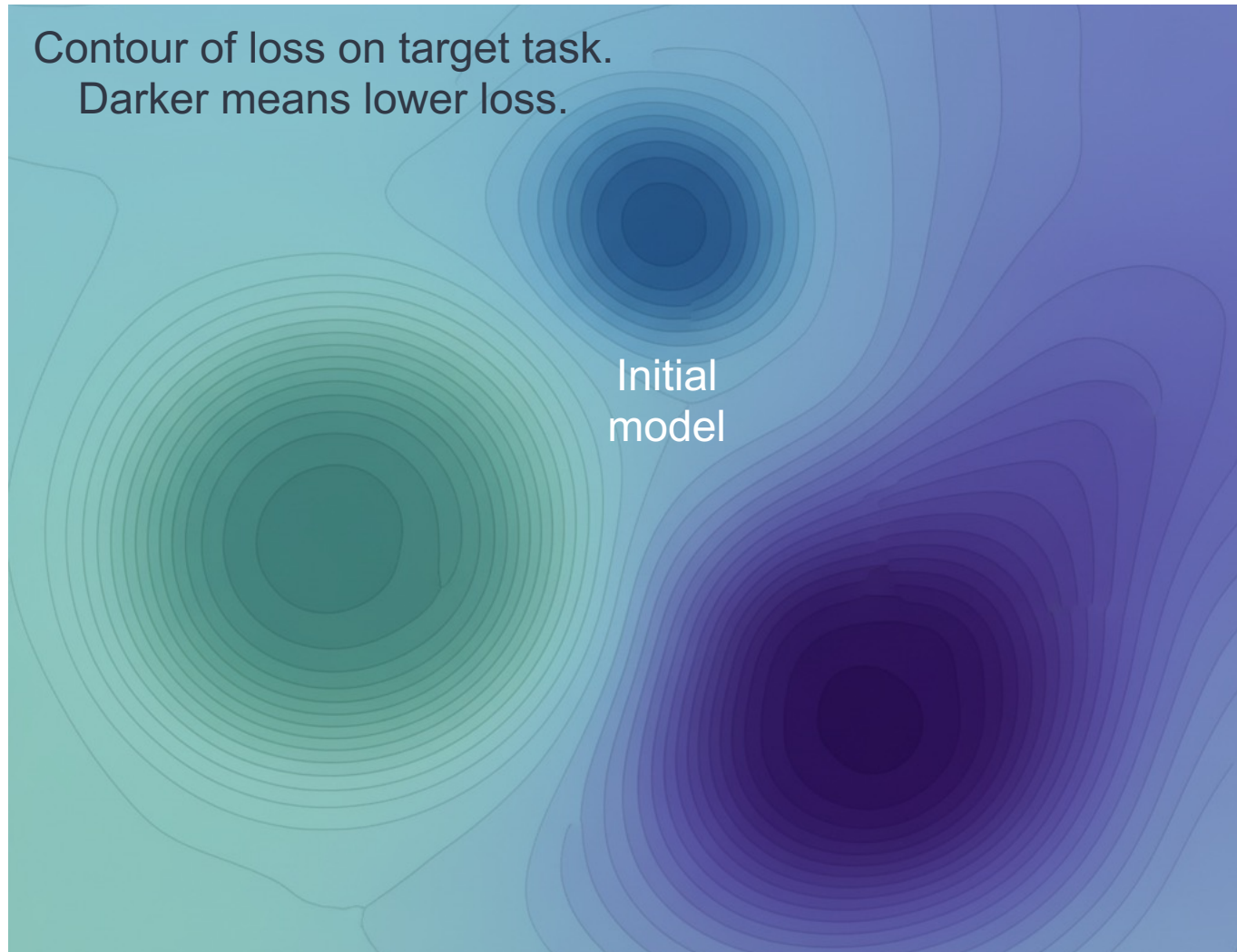


What (we think) IA2 is doing?

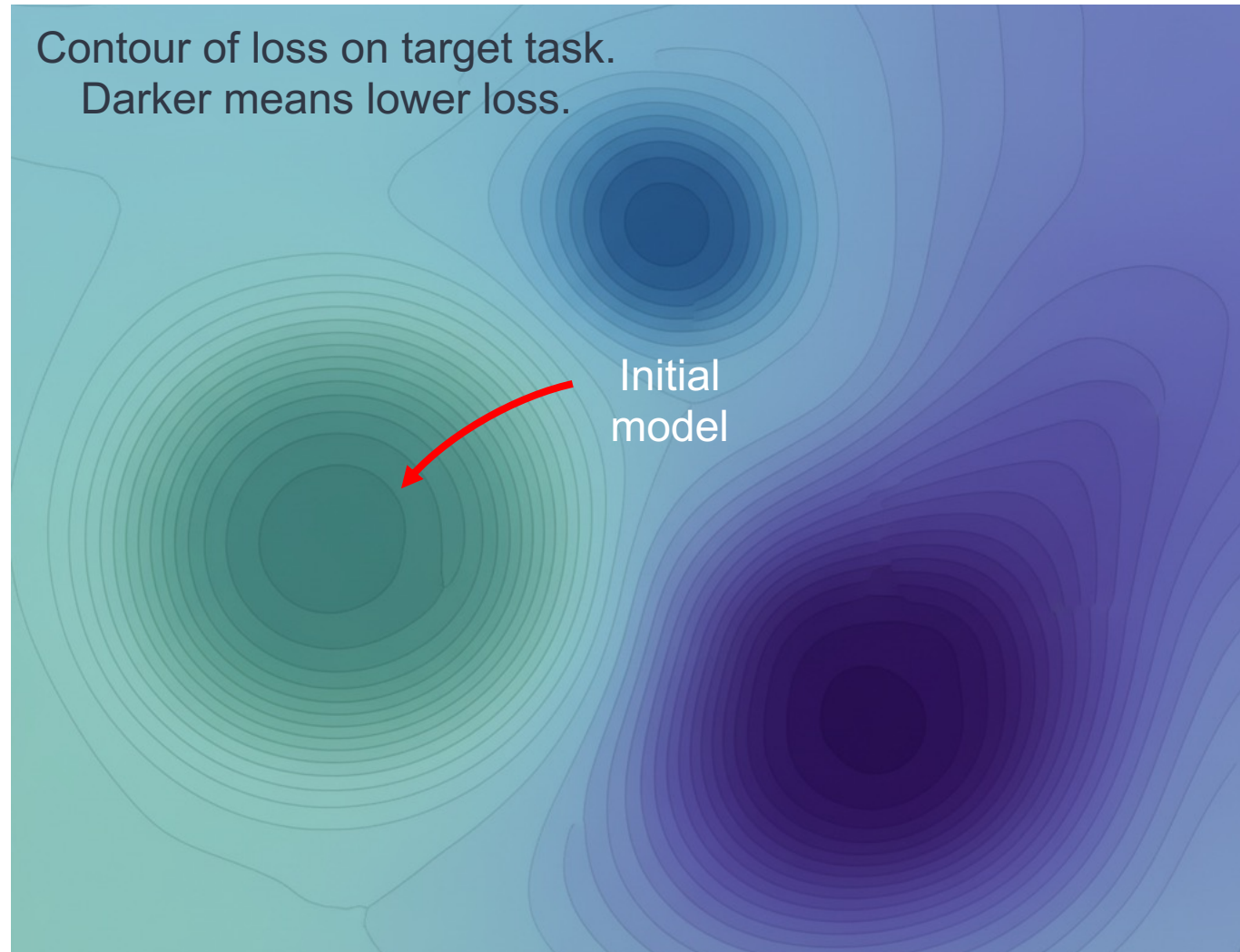
Contour of loss on target task.
Darker means lower loss.



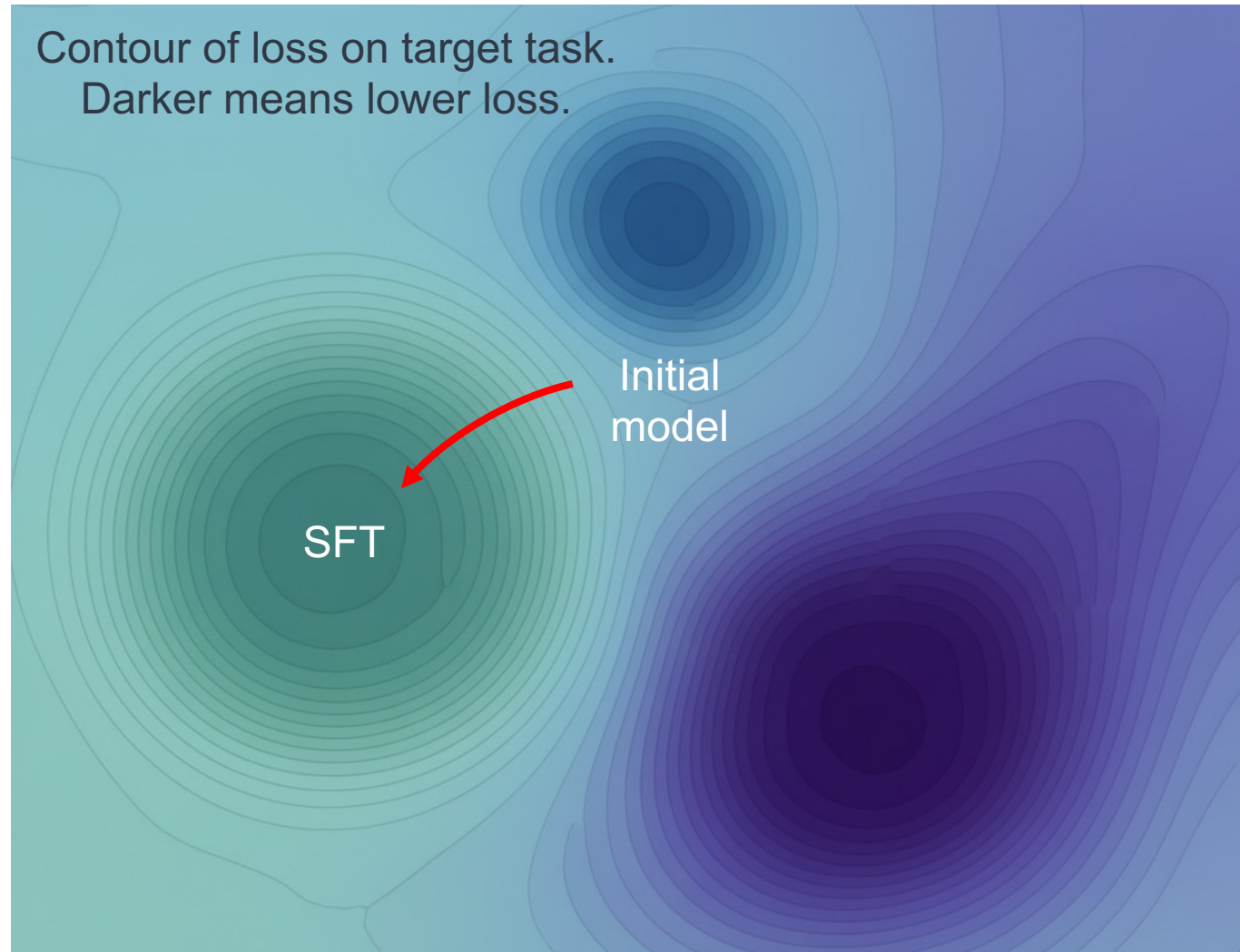
What (we think) IA2 is doing?



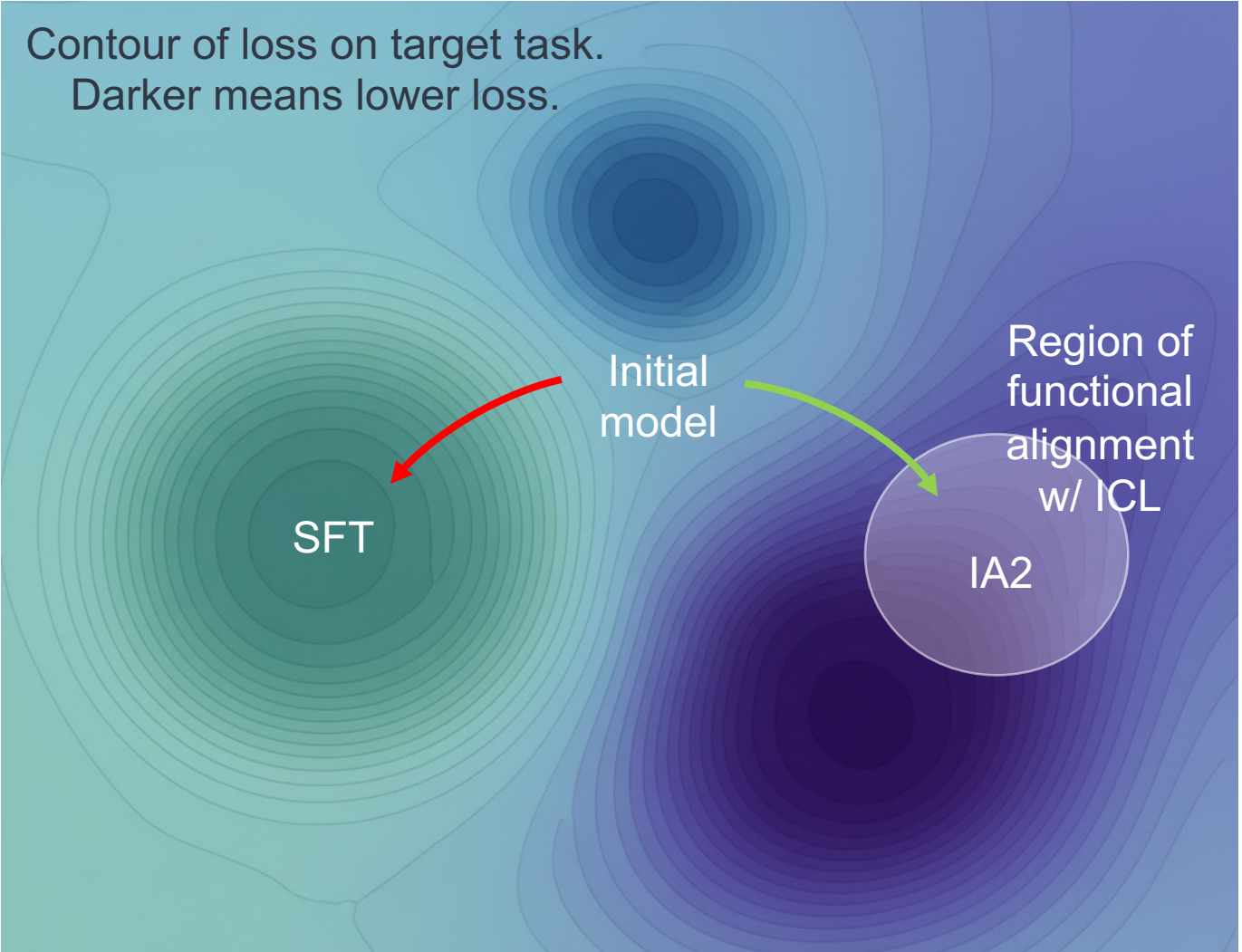
What (we think) IA2 is doing?



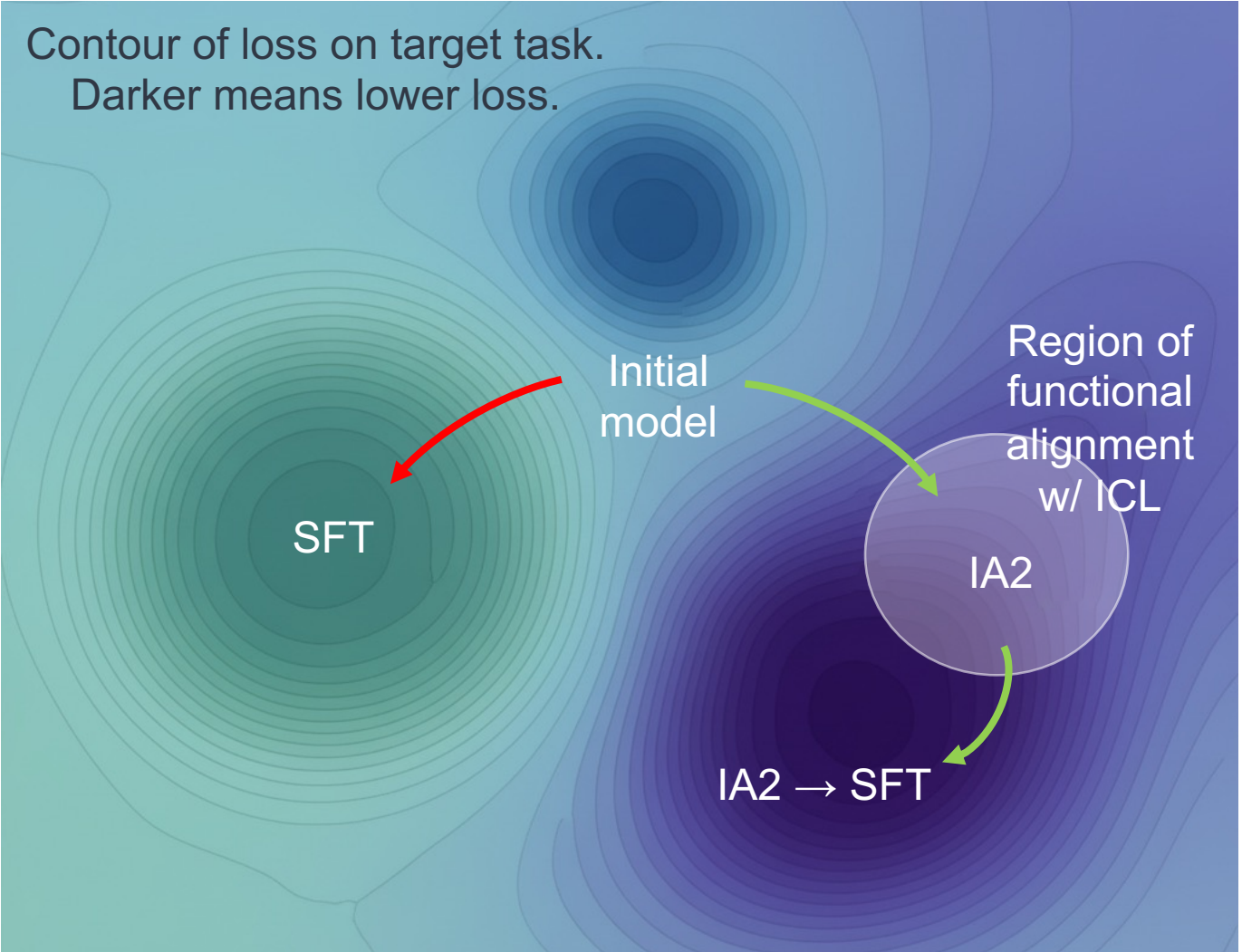
What (we think) IA2 is doing?



What (we think) IA2 is doing?



What (we think) IA2 is doing?



Summary

- **SFT and ICL differ internally** — not just in outputs, but in how they process inputs at every layer.
- **IA2 bridges this gap** — a lightweight priming step that aligns SFT weights to ICL's activation patterns.
- **Empirical results:** Consistent gains over SFT-only.

Summary

- **SFT and ICL differ internally** — not just in outputs, but in how they process inputs at every layer.
- **IA2 bridges this gap** — a lightweight priming step that aligns SFT weights to ICL's activation patterns.
- **Empirical results:** Consistent gains over SFT-only.

Summary

- **SFT and ICL differ internally** — not just in outputs, but in how they process inputs at every layer.
- **IA2 bridges this gap** — a lightweight priming step that aligns SFT weights to ICL's activation patterns.
- **Empirical results:** Consistent gains over SFT-only.

Summary

- **SFT and ICL differ internally** — not just in outputs, but in how they process inputs at every layer.
- **IA2 bridges this gap** — a lightweight priming step that aligns SFT weights to ICL's activation patterns.
- **Empirical results:** Consistent gains over SFT-only.

Open questions

- Lots of recent works on “self-distillation”
 - all use SFT-like objectives (no representation-level objectives)

Reinforcement Learning via Self-Distillation

Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models

Siyao Zhao^{†1} Zhihui Xie² Mengchen Liu³ Jing Huang³ Guan Pang³ Feiyu Chen^{*,†3} Aditya Grover^{*1}

Robert Ge¹ Frederike Lübeck^{*,1,2} Lejs Behrlic^{*,1} Anton Baumann^{*,1}
Guglielmo Agatella^{1,2} Daniel Marta¹ Ido Hakimi¹ Idan Shenfeld³
Kleine Buening¹ Carlos Guestrin⁴ Andreas Krause¹
Thomas Roth¹ ²Max Planck Institute for Intelligent Systems ³MIT ⁴Stanford

- **Open question:** representation-level distillation of long reasoning chains.

Open questions

- Lots of recent works on “self-distillation”
 - all use SFT-like objectives (no representation-level objectives)

Reinforcement Learning via Self-Distillation

Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models

Siyao Zhao^{†1} Zhihui Xie² Mengchen Liu³ Jing Huang³ Guan Pang³ Feiyu Chen^{*,‡3} Aditya Grover^{*1}

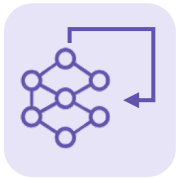
Robert Ge¹ Frederike Lübeck^{*,1,2} Lejs Behrić^{*,1} Anton Baumann^{*,1}
Guglielmo Agatella^{1,2} Daniel Marta¹ Ido Hakimi¹ Idan Shenfeld³
Kleine Buening¹ Carlos Guestrin⁴ Andreas Krause¹
Thomas Roth¹ ²Max Planck Institute for Intelligent Systems ³MIT ⁴Stanford

- **Open question:** representation-level distillation of long reasoning chains.

Today: Rethinking the Reasoning Stack



1. How does in-context learning (ICL) **emerge**? (pre-training)



2. How do we **distill ICL**? (pre → post)

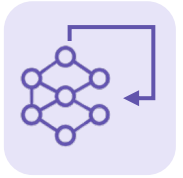


3. How do we maintain **diversity**? (post-training)

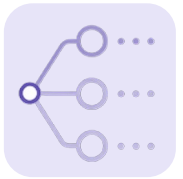
Today: Rethinking the Reasoning Stack



1. How does in-context learning (ICL) **emerge**? (pre-training)



2. How do we **distill ICL**? (pre \rightarrow post)



3. How do we maintain **diversity**? (post-training)

pick a random animal

How about a axolotl?

pick a random animal



How about a axolotl?

pick a random animal



How about a axolotl?

pick a random animal

And I asked it
10 times!



How about a axolotl?

8x



Let's go with: the Quokka 🐾

1x



Let's go with the Fennec fox 🦊

1x

pick a random animal

And I asked it
10 times!

LLMs often generate repetitive responses



How about a axolotl?

8x



Let's go with: the Quokka 🐾

1x



Let's go with the Fennec fox 🦊

1x

pick a random animal

And I asked it
10 times!

Why care about “diversity”?

Why care about “diversity”?

- **What is creativity?** Creativity is the production of outputs that are both *novel* and *useful*.
- **What enables “novelty”?** Being able to explore *a wide space of possible reasons/arguments*.
- **Therefore: diversity is the prerequisite:** Without diversity of outputs, there is no novelty.

Why care about “diversity”?

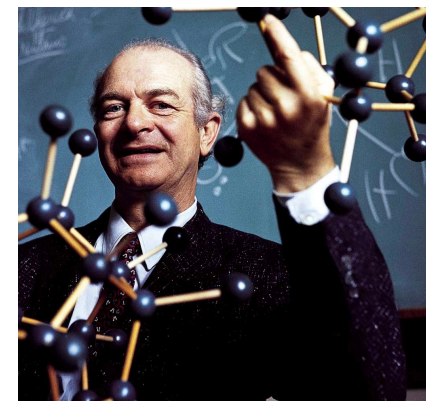
- **What is creativity?** Creativity is the production of outputs that are both *novel* and *useful*.
- **What enables “novelty”?** Being able to explore *a wide space of possible reasons/arguments*.
- **Therefore: diversity is the prerequisite:** Without diversity of outputs, there is no novelty.

Why care about “diversity”?

- **What is creativity?** Creativity is the production of outputs that are both *novel* and *useful*.
- **What enables “novelty”?** Being able to explore *a wide space of possible reasons/arguments*.
- **Therefore: diversity is the prerequisite:** Without diversity of outputs, there is no novelty.

Lack of diversity \Rightarrow lack of “creative” reasoning

“The best way to have a good idea is to have lots of ideas.” –Linus Pauling (1901-1994)



Building toward a solution

- We need a *sequence-level* diversity signal
 - *Token-level* interventions tend to be brittle.
 - *Temperature scaling* (during sampling) trades off quality for diversity. [Peeperkorn et al., 2024]
 - *Entropy/unlikelihood regularization* (during training) don't necessarily lead to sequence-level diversity. [Cheng et al., 2025; He et al., 2025]
- Need *online* algorithm to allow the diversity signal to adapt as the model evolves during training.
 - *Offline* approaches don't adapt to the model. [Lanchantin et al., 2025]

Building toward a solution

- We need a *sequence-level* diversity signal
 - *Token-level* interventions tend to be brittle.
 - *Temperature scaling* (during sampling) trades off quality for diversity. [Peeperkorn et al., 2024]
 - *Entropy/unlikelihood regularization* (during training) don't necessarily lead to sequence-level diversity. [Cheng et al., 2025; He et al., 2025]
- Need *online* algorithm to allow the diversity signal to adapt as the model evolves during training.
 - *Offline* approaches don't adapt to the model. [Lanchantin et al., 2025]

Building toward a solution

- We need a *sequence-level* diversity signal
 - **Token-level** interventions tend to be brittle.
 - **Temperature scaling** (during sampling) trades off quality for diversity. [Peeperkorn et al., 2024]
 - **Entropy/unlikelihood regularization** (during training) don't necessarily lead to sequence-level diversity. [Cheng et al., 2025; He et al., 2025]
- Need *online* algorithm to allow the diversity signal to adapt as the model evolves during training.
 - *Offline* approaches don't adapt to the model. [Lanchantin et al., 2025]

Building toward a solution

- We need a *sequence-level* diversity signal
 - **Token-level** interventions tend to be brittle.
 - **Temperature scaling** (during sampling) trades off quality for diversity. [Peeperkorn et al., 2024]
 - **Entropy/unlikelihood regularization** (during training) don't necessarily lead to sequence-level diversity. [Cheng et al., 2025; He et al., 2025]
- Need *online* algorithm to allow the diversity signal to adapt as the model evolves during training.
 - *Offline* approaches don't adapt to the model. [Lanchantin et al., 2025]

Building toward a solution

- We need a *sequence-level* diversity signal
 - **Token-level** interventions tend to be brittle.
 - **Temperature scaling** (during sampling) trades off quality for diversity. [Peeperkorn et al., 2024]
 - **Entropy/unlikelihood regularization** (during training) don't necessarily lead to sequence-level diversity. [Cheng et al., 2025; He et al., 2025]
- Need *online* algorithm to allow the diversity signal to adapt as the model evolves during training.
 - *Offline* approaches don't adapt to the model. [Lanchantin et al., 2025]

Building toward a solution

- We need a *sequence-level* diversity signal
 - **Token-level** interventions tend to be brittle.
 - **Temperature scaling** (during sampling) trades off quality for diversity. [Peeperkorn et al., 2024]
 - **Entropy/unlikelihood regularization** (during training) don't necessarily lead to sequence-level diversity. [Cheng et al., 2025; He et al., 2025]
- Need *online* algorithm to allow the diversity signal to adapt as the model evolves during training.
 - *Offline* approaches don't adapt to the model. [Lanchantin et al., 2025]

Building toward a solution

- We need a *sequence-level* diversity signal
 - **Token-level** interventions tend to be brittle.
 - **Temperature scaling** (during sampling) trades off quality for diversity. [Peeperkorn et al., 2024]
 - **Entropy/unlikelihood regularization** (during training) don't necessarily lead to sequence-level diversity. [Cheng et al., 2025; He et al., 2025]
- Need *online* algorithm to allow the diversity signal to adapt as the model evolves during training.
 - **Offline** approaches don't adapt to the model. [Lanchantin et al., 2025]

GRPO-style RL

GRPO-style RL

Input

Write a story
about a
programmer with
superpowers.

GRPO-style RL

Input

Write a story
about a
programmer with
superpowers.



LM

GRPO-style RL

Input

Write a story
about a
programmer with
superpowers.

LM

Rollouts

(a) Lena woke up and reached to
her PC 🖥️. She realized that she
can debug without even looking
at the code...

GRPO-style RL

Input

Write a story about a programmer with superpowers.

LM

Rollouts

(a) Lena woke up and reached to her PC 🖥️. She realized that she can debug without even looking at the code...

(b) Eli used to be an ordinary backend. Suddenly his code began to debug themselves...

GRPO-style RL

Input

Write a story about a programmer with superpowers.

LM

Rollouts

(a) Lena woke up and reached to her PC `█`. She realized that she can debug without even looking at the code...

(b) Eli used to be an ordinary backend. Suddenly his code began to debug themselves...

(c) Nira found a note. She simply wrote `height+=1` on it, and she grew one inch taller! She then wrote `bank_balance*=100`...

GRPO-style RL

Input

Write a story about a programmer with superpowers.

LM

Rollouts

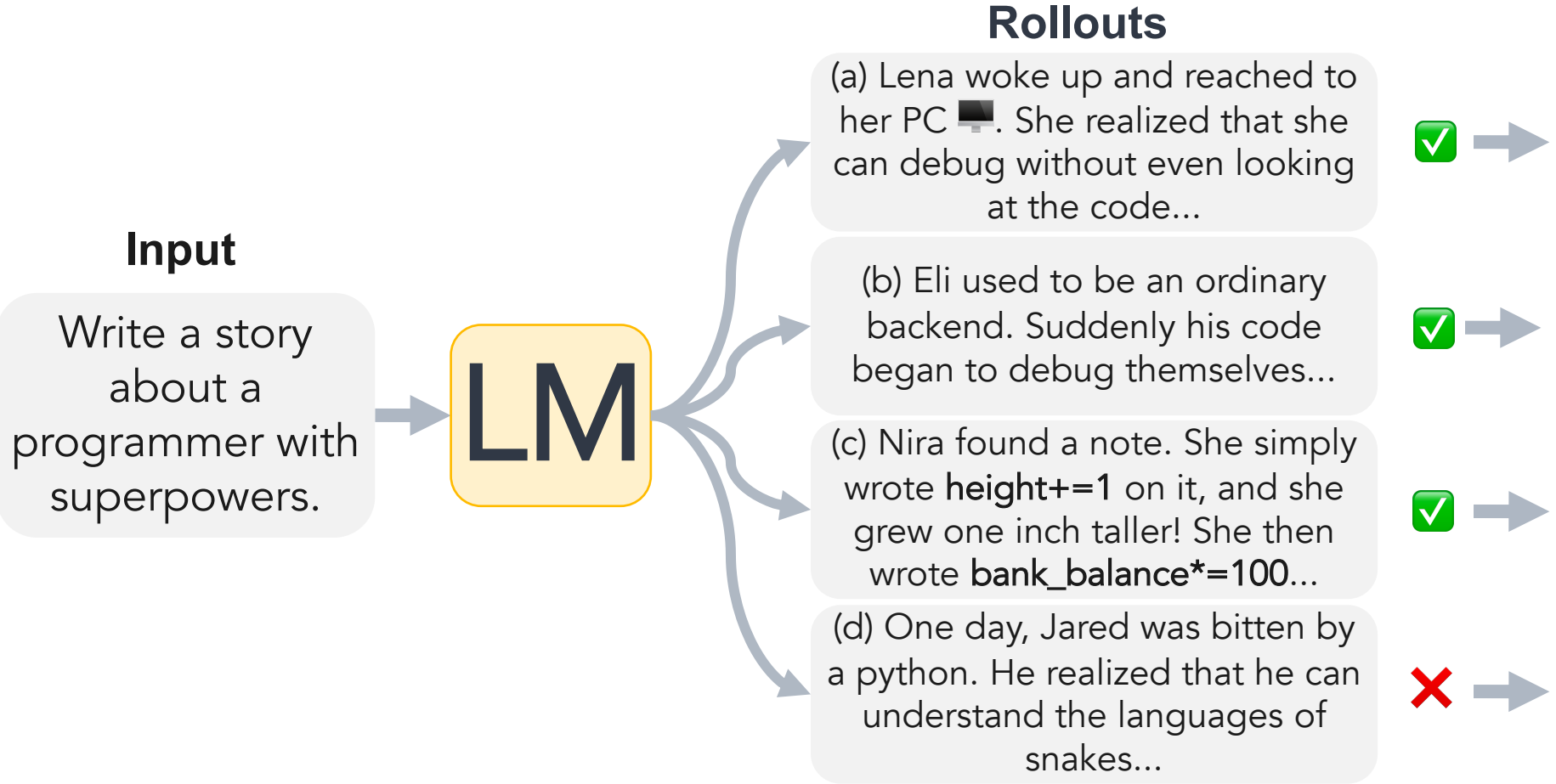
(a) Lena woke up and reached to her PC `█`. She realized that she can debug without even looking at the code...

(b) Eli used to be an ordinary backend. Suddenly his code began to debug themselves...

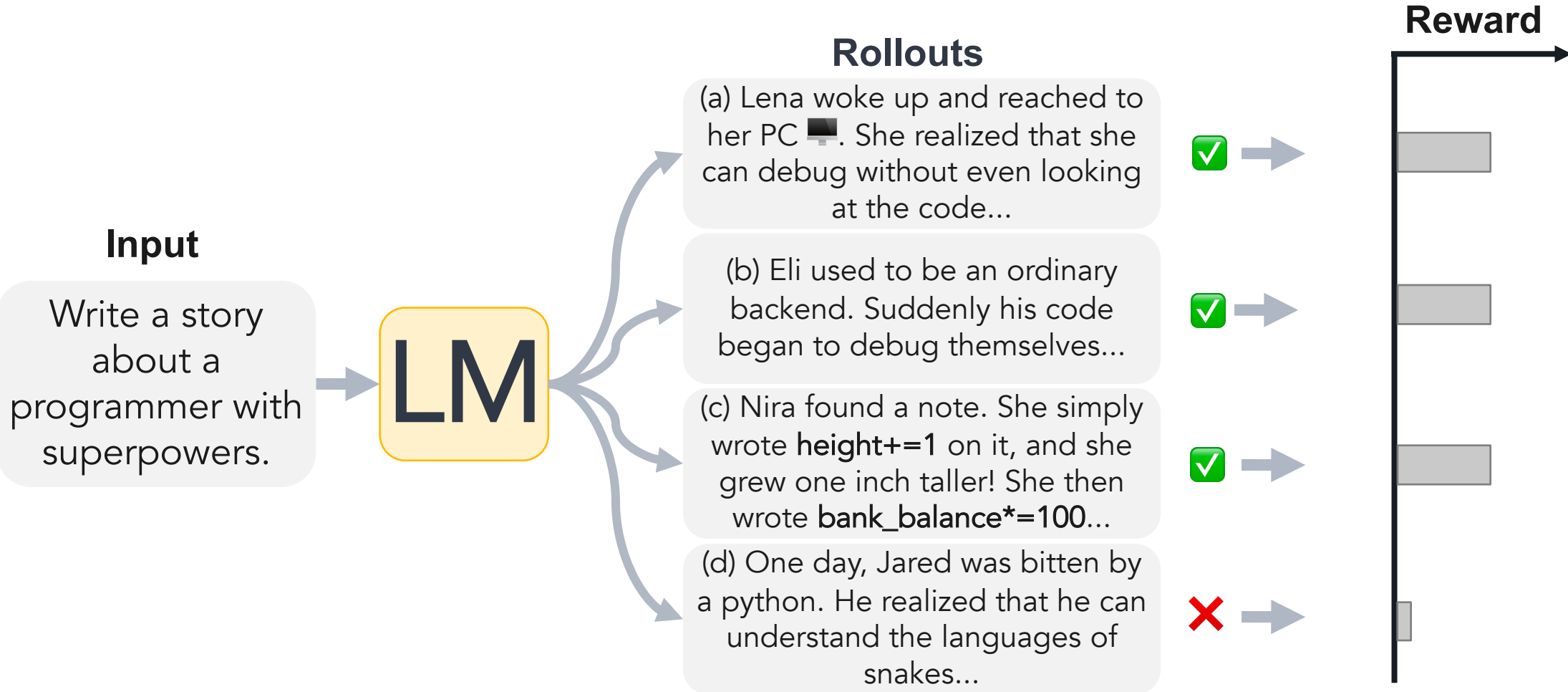
(c) Nira found a note. She simply wrote `height+=1` on it, and she grew one inch taller! She then wrote `bank_balance*=100`...

(d) One day, Jared was bitten by a python. He realized that he can understand the languages of snakes...

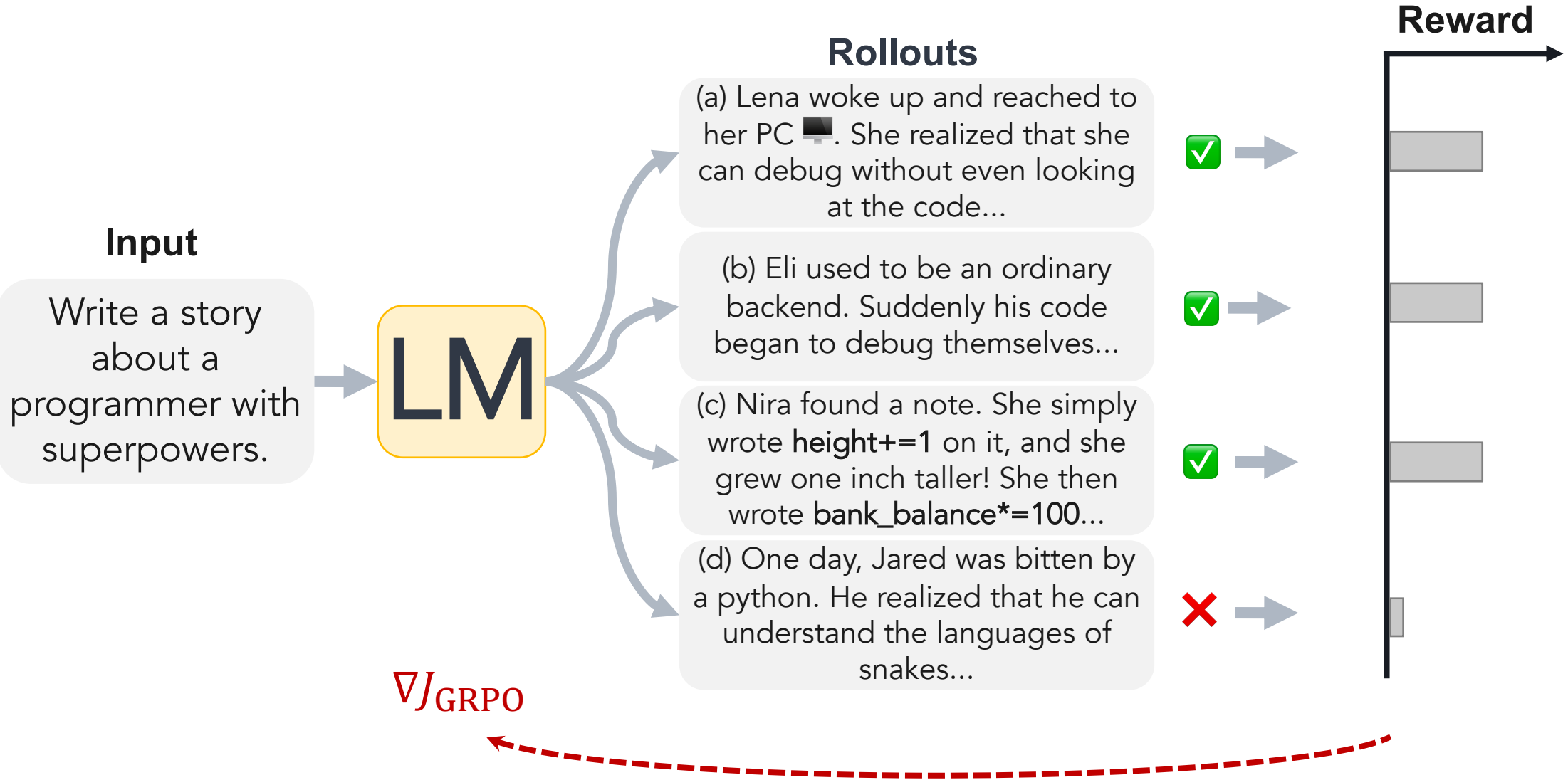
GRPO-style RL



GRPO-style RL



GRPO-style RL



GRPO-style RL

Doesn't consider whether different responses are equivalent.

Input

Write a story about a programmer with superpowers.

LM

Rollouts

(a) Lena woke up and reached to her PC `█`. She realized that she can debug without even looking at the code...



(b) Eli used to be an ordinary backend. Suddenly his code began to debug themselves...



(c) Nira found a note. She simply wrote `height+=1` on it, and she grew one inch taller! She then wrote `bank_balance*=100`...



(d) One day, Jared was bitten by a python. He realized that he can understand the languages of snakes...



Reward

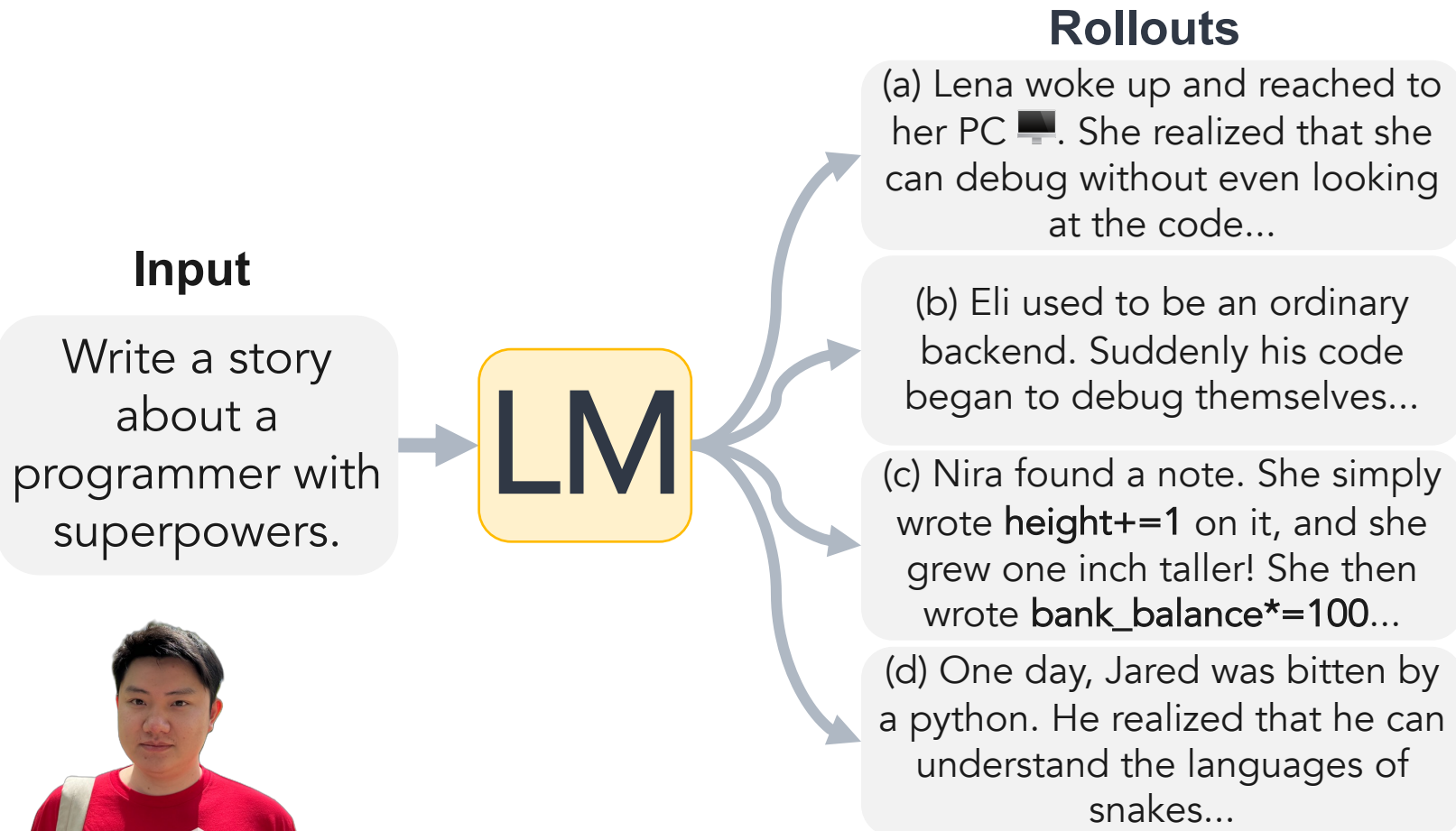


∇J_{GRPO}



Diversity-aware Reinforcement Learning (DARLING)

- **Main idea:** Reward responses for both quality **and diversity**.



Diversity-aware Reinforcement Learning (DARLING)

- **Step 1:** Determine whether two responses are “similar”.

Rollouts

Input

Write a story about a programmer with superpowers.

LM

(a) Lena woke up and reached to her PC `█`. She realized that she can debug without even looking at the code...

(b) Eli used to be an ordinary backend. Suddenly his code began to debug themselves...

(c) Nira found a note. She simply wrote `height+=1` on it, and she grew one inch taller! She then wrote `bank_balance*=100`...

(d) One day, Jared was bitten by a python. He realized that he can understand the languages of snakes...

Diversity-aware Reinforcement Learning (DARLING)

- **Step 1:** Determine whether two responses are “similar”.

Rollouts

(a) Lena woke up and reached to her PC `█`. She realized that she can debug without even looking at the code...

(b) Eli used to be an ordinary backend. Suddenly his code began to debug themselves...

(c) Nira found a note. She simply wrote `height+=1` on it, and she grew one inch taller! She then wrote `bank_balance*=100`...

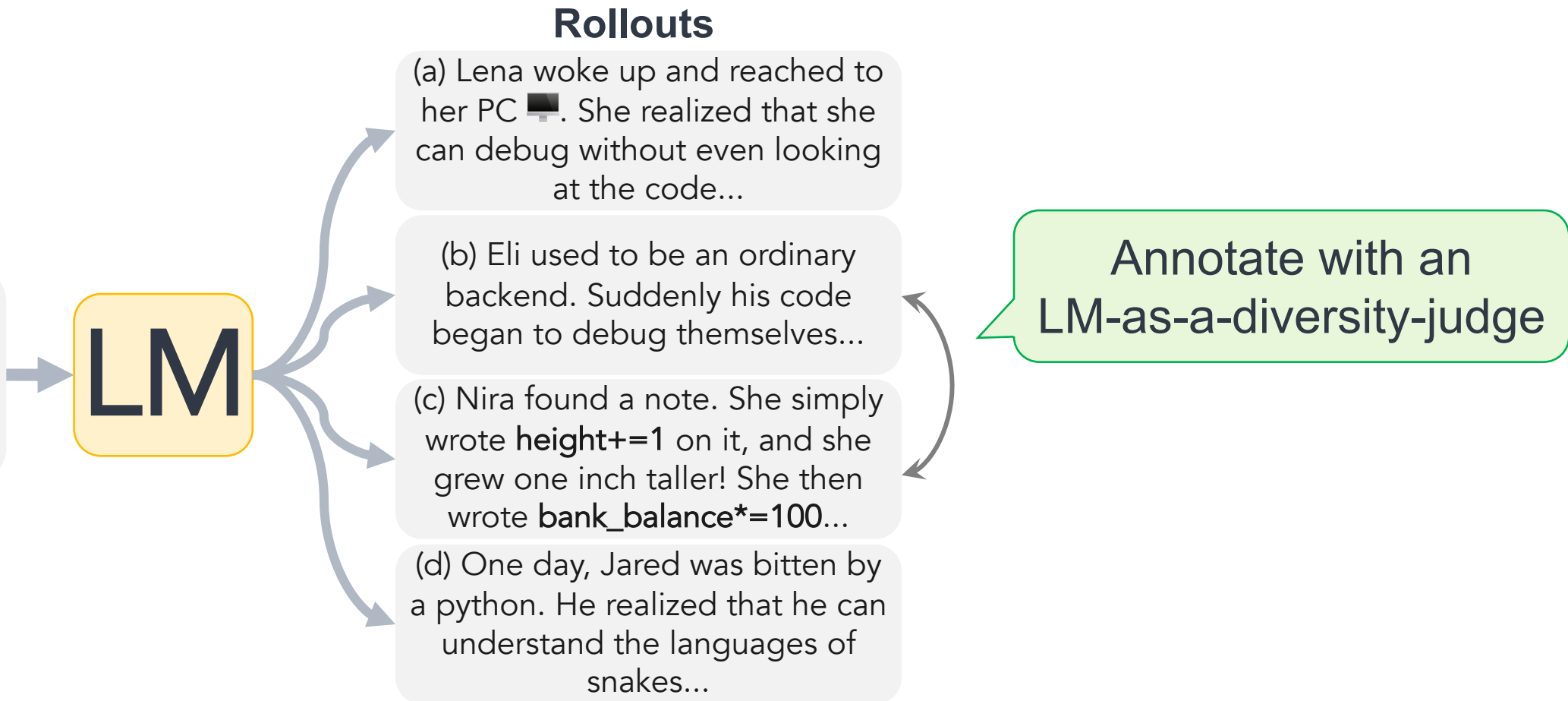
(d) One day, Jared was bitten by a python. He realized that he can understand the languages of snakes...

Annotate with an LM-as-a-diversity-judge

Input

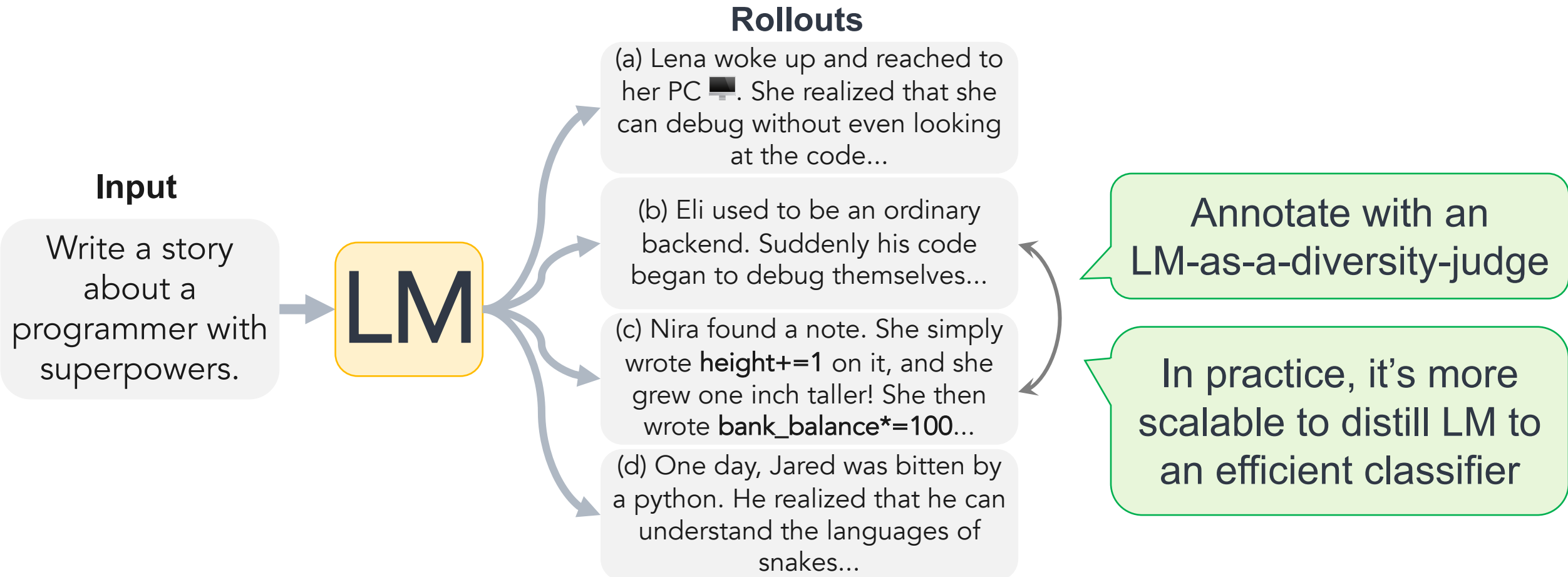
Write a story about a programmer with superpowers.

LM



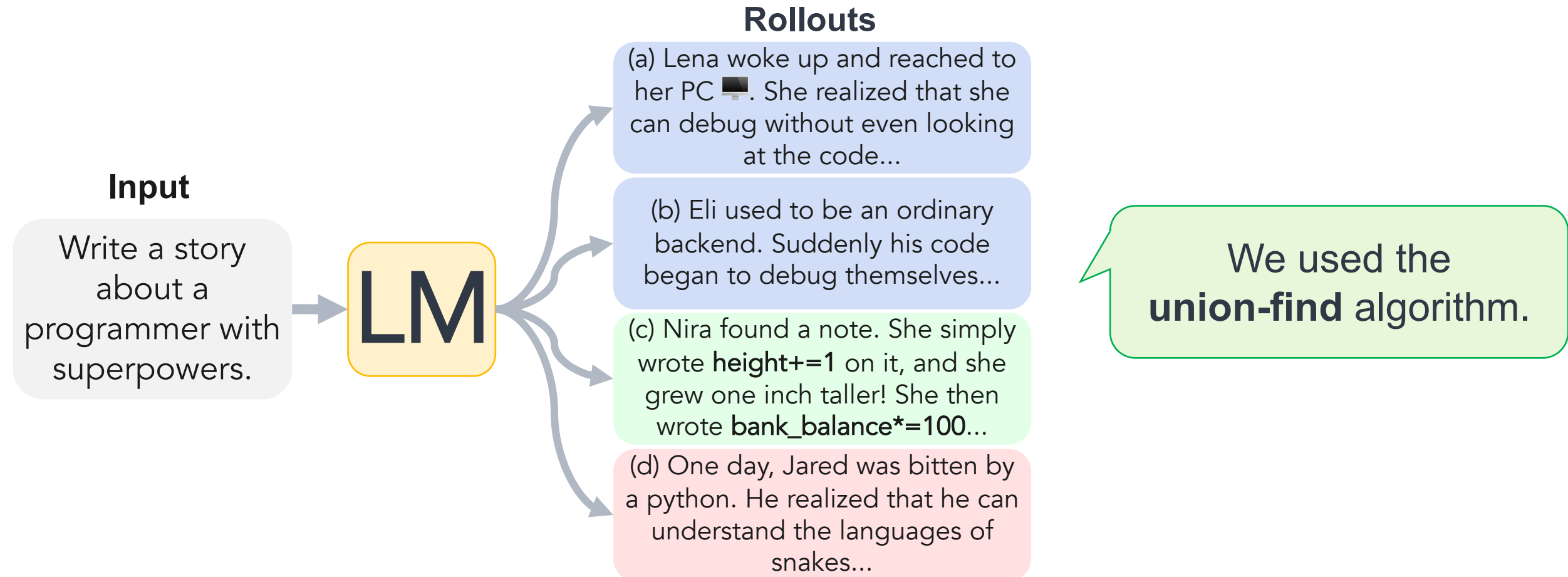
Diversity-aware Reinforcement Learning (DARLING)

- **Step 1:** Determine whether two responses are “similar”.



Diversity-aware Reinforcement Learning (DARLING)

- **Step 2:** Partition the rollouts into semantic equivalence classes.



Diversity-aware Reinforcement Learning (DARLING)


- **Step 3:** Assign “diversity rewards” := $\frac{\text{\# of responses not in the same group}}{\text{Total \# of responses} - 1}$

Rollouts

Input

Write a story about a programmer with superpowers.

LM

(a) Lena woke up and reached to her PC . She realized that she can debug without even looking at the code...

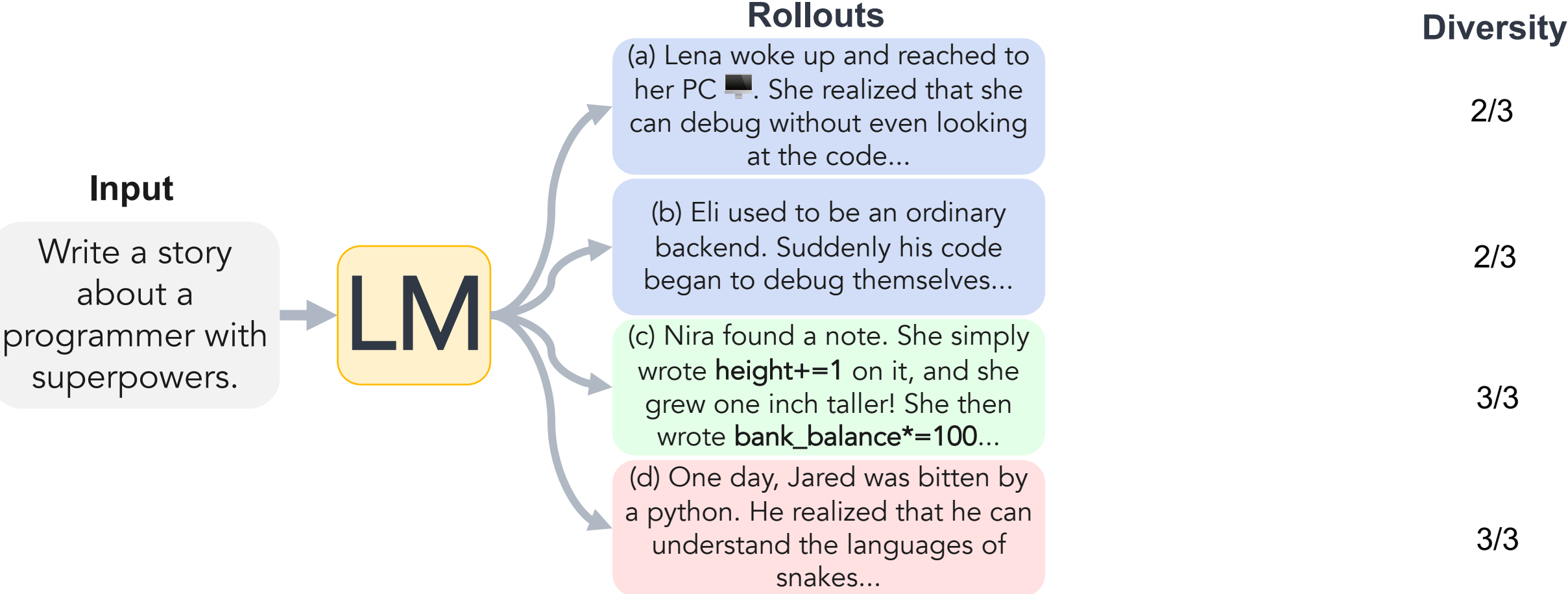
(b) Eli used to be an ordinary backend. Suddenly his code began to debug themselves...

(c) Nira found a note. She simply wrote `height+=1` on it, and she grew one inch taller! She then wrote `bank_balance*=100`...

(d) One day, Jared was bitten by a python. He realized that he can understand the languages of snakes...

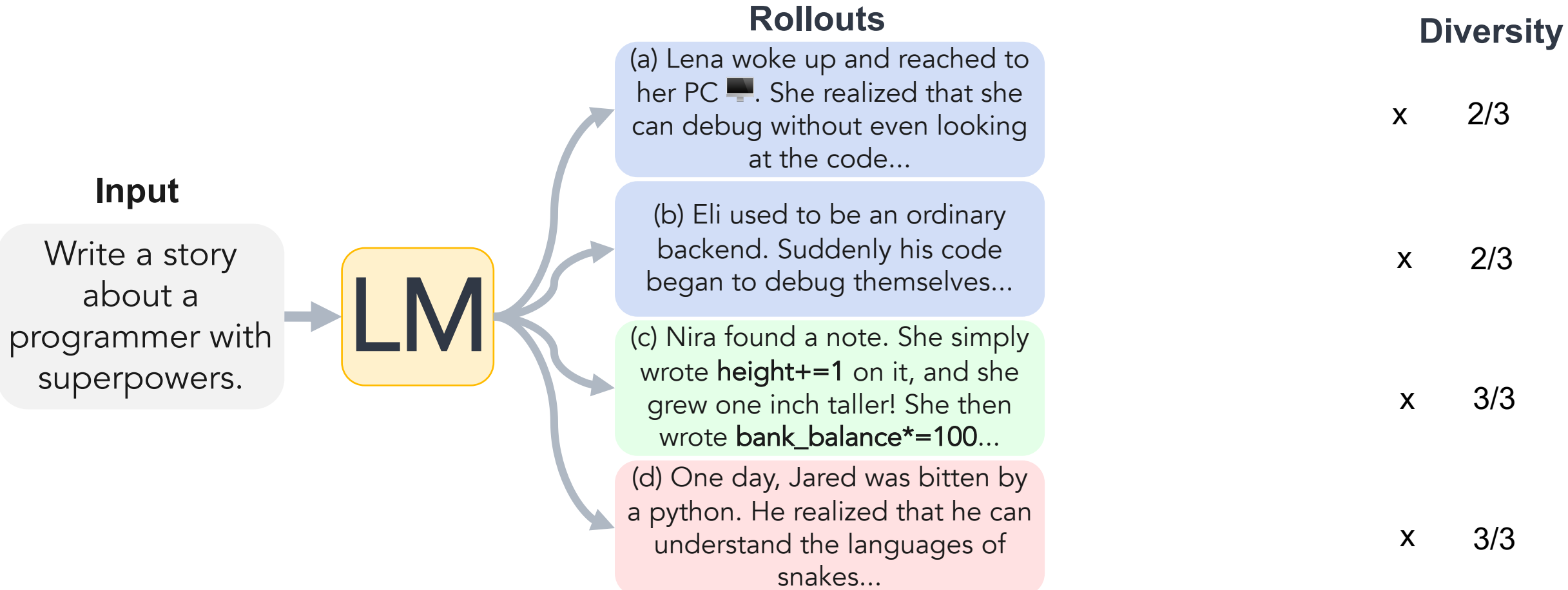
Diversity-aware Reinforcement Learning (DARLING)

- **Step 3:** Assign “diversity rewards” := $\frac{\text{\# of responses not in the same group}}{\text{Total \# of responses} - 1}$



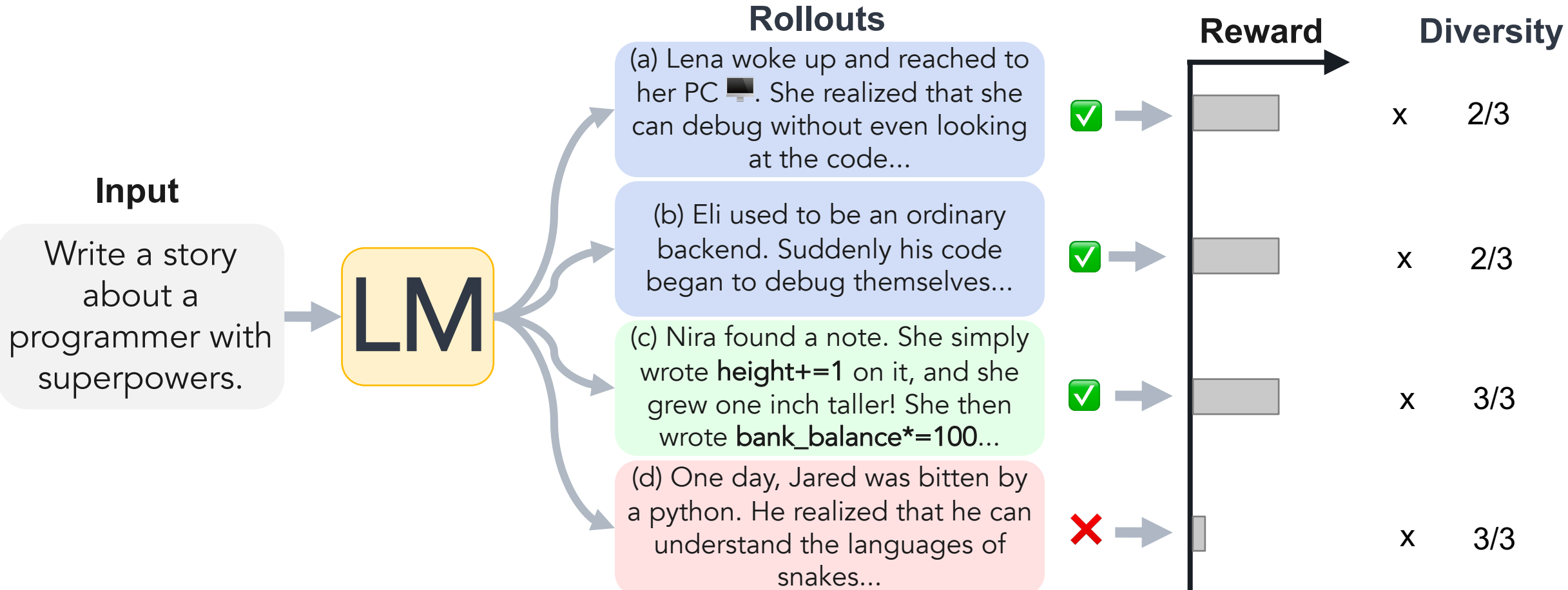
Diversity-aware Reinforcement Learning (DARLING)

• **Step 3:** Assign “diversity rewards” := $\frac{\text{\# of responses not in the same group}}{\text{Total \# of responses} - 1}$



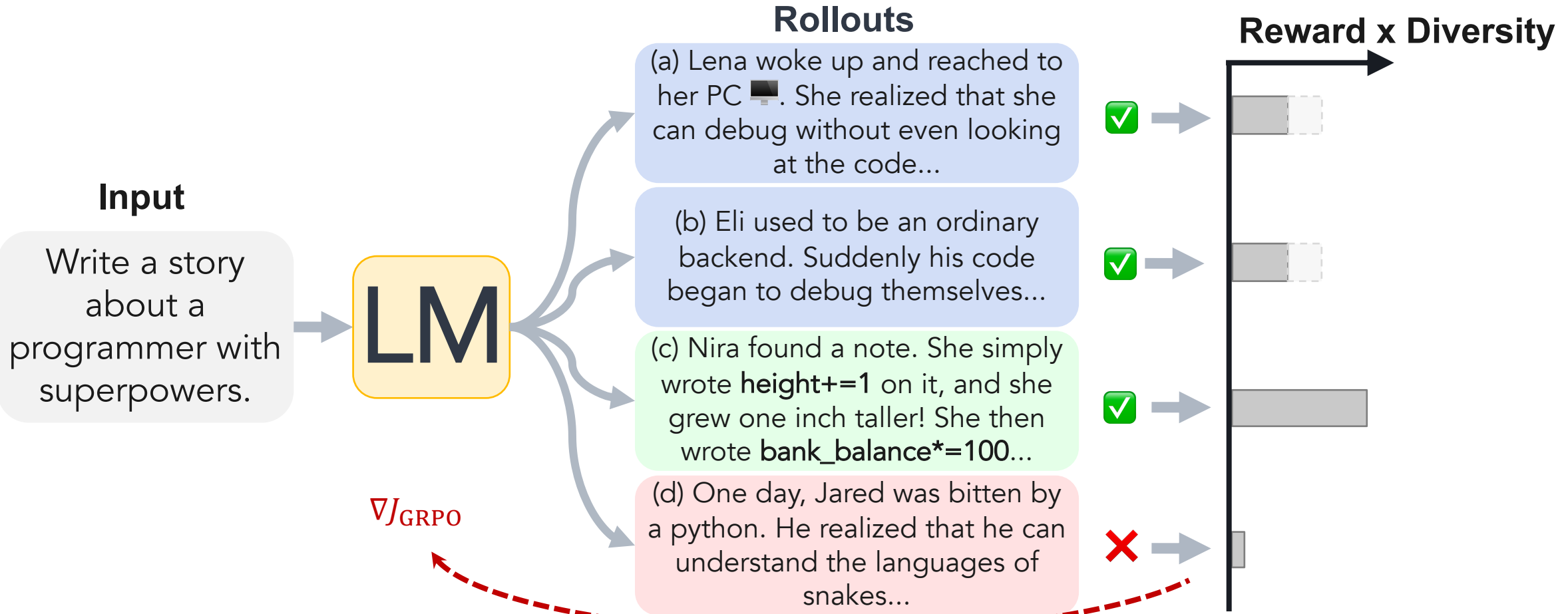
Diversity-aware Reinforcement Learning (DARLING)

• **Step 3:** Assign “diversity rewards” := $\frac{\text{\# of responses not in the same group}}{\text{Total \# of responses} - 1}$

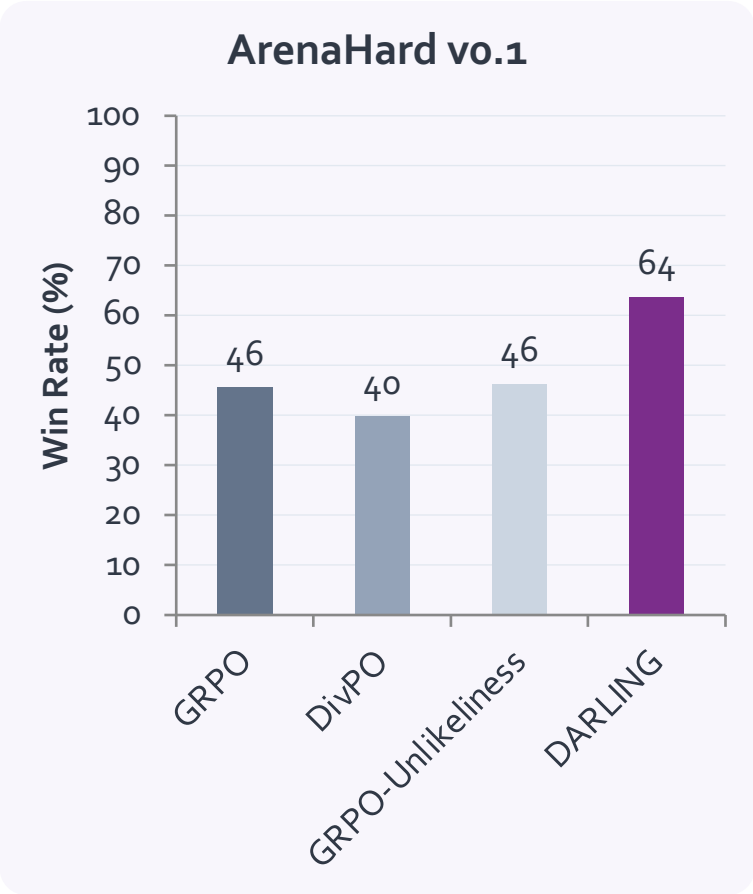
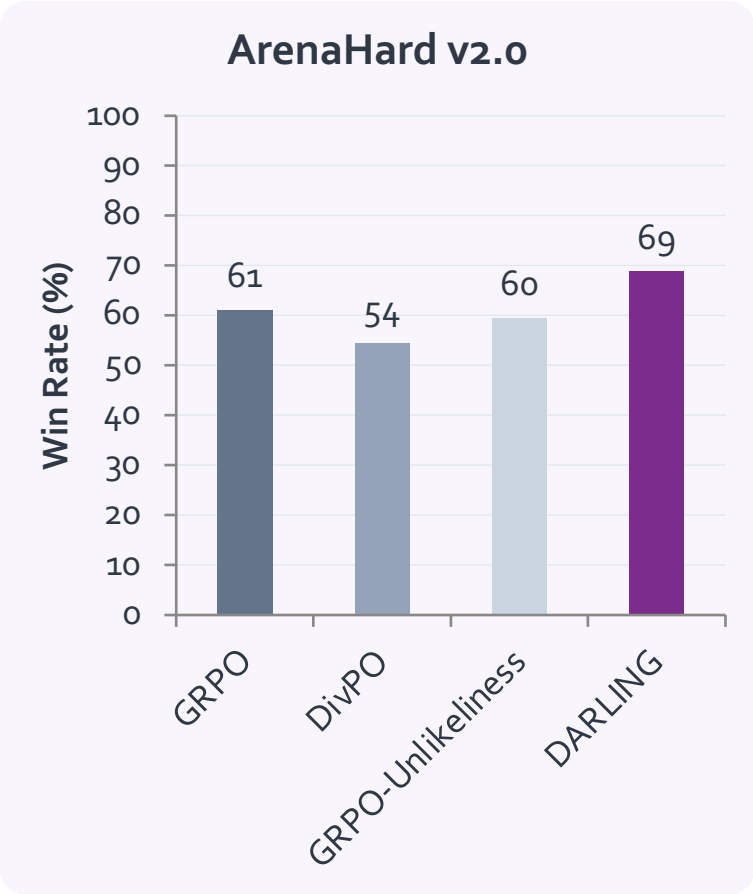
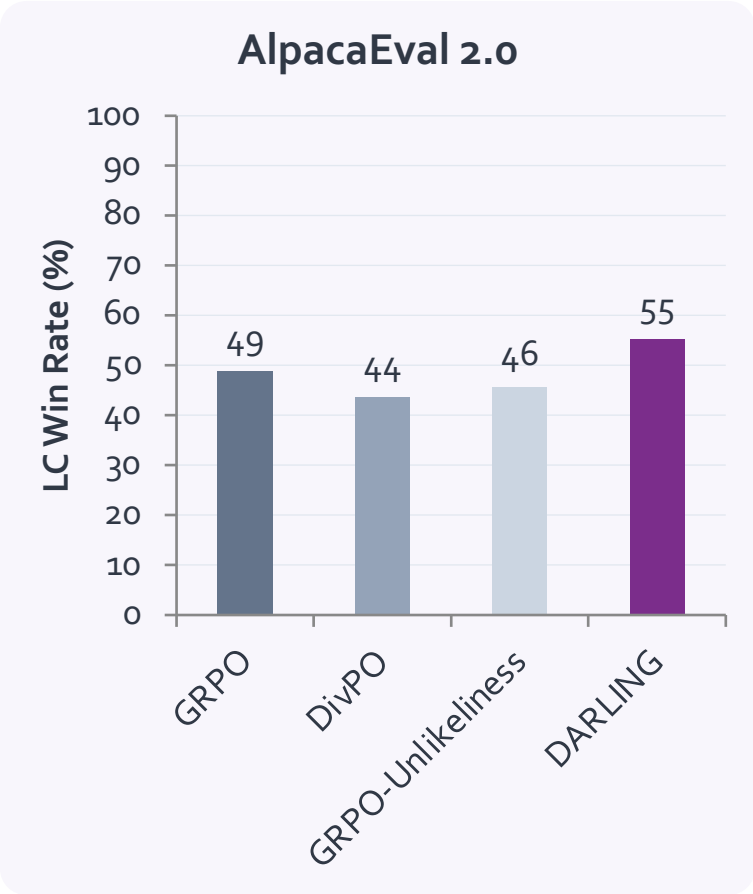


Diversity-aware Reinforcement Learning (DARLING)

- **Step 3:** Assign “diversity rewards” := $\frac{\text{\# of responses not in the same group}}{\text{Total \# of responses} - 1}$



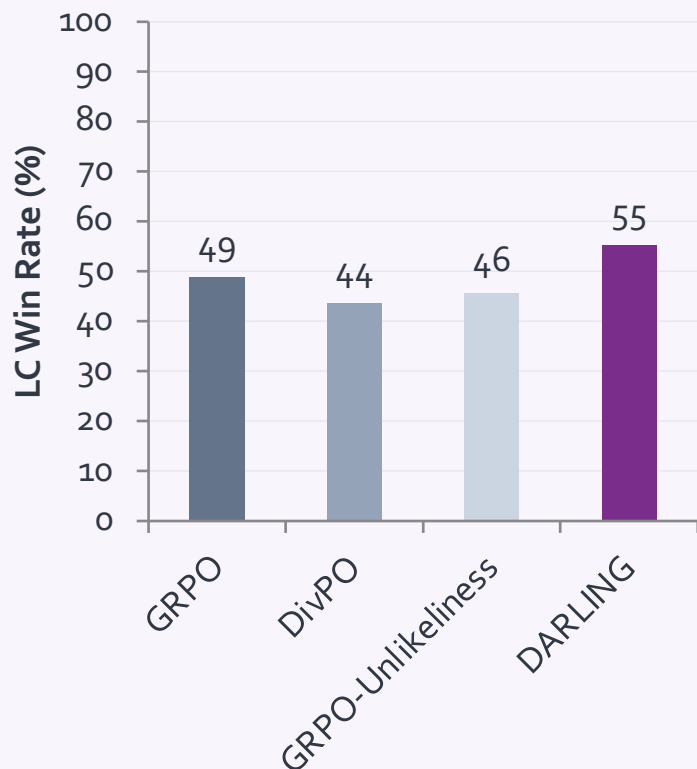
Empirical Results



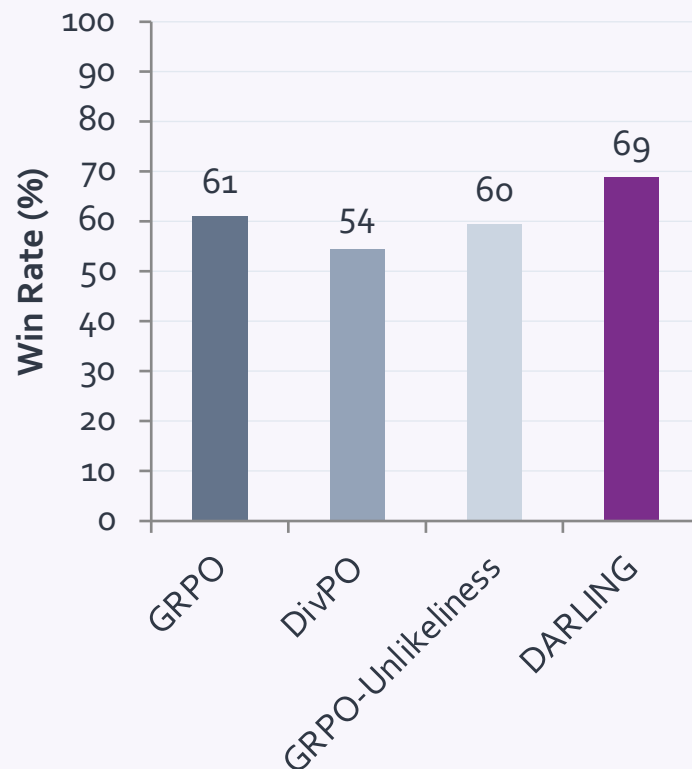
Llama-3.1-8B-Instruct

DARLING improves quality ...

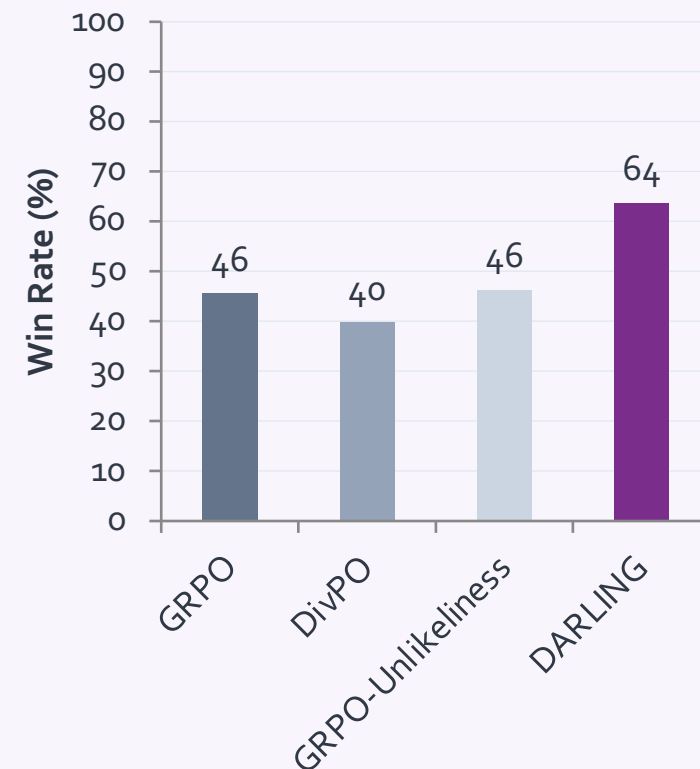
AlpacaEval 2.0



ArenaHard v2.0



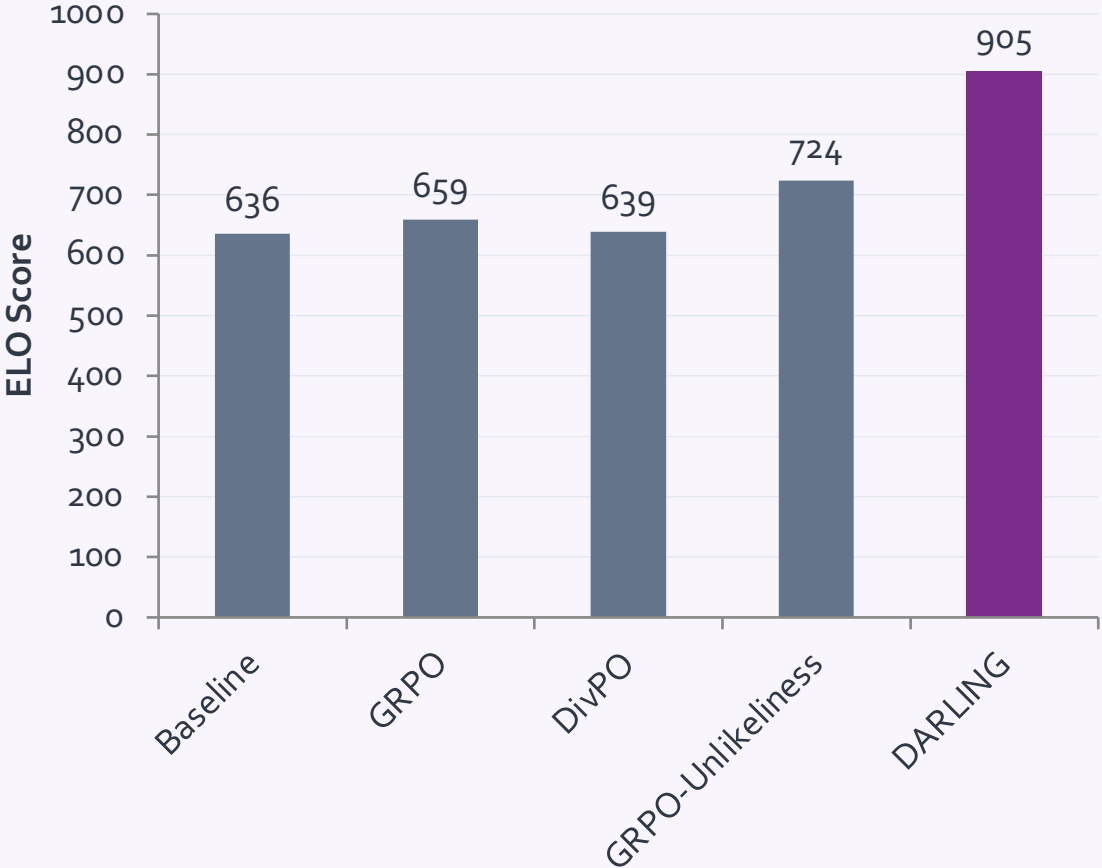
ArenaHard vo.1



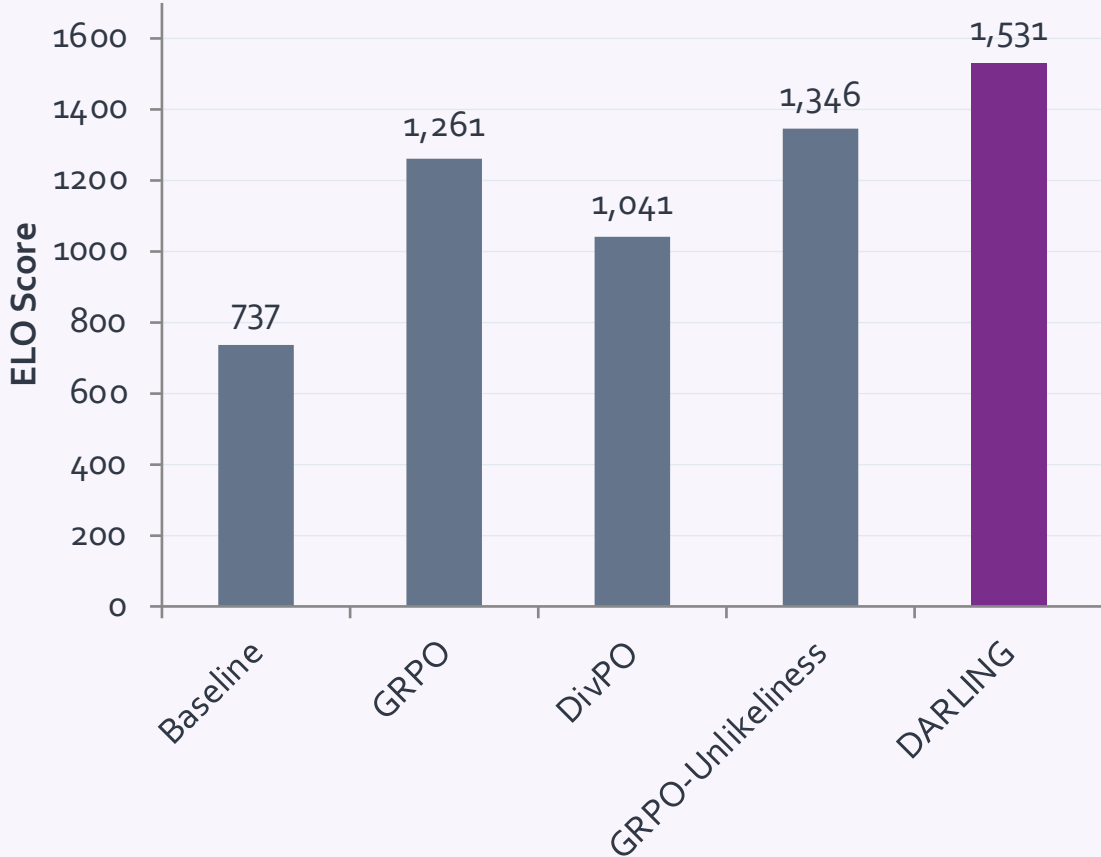
Llama-3.1-8B-Instruct

Empirical Results

Llama-3.1-8B



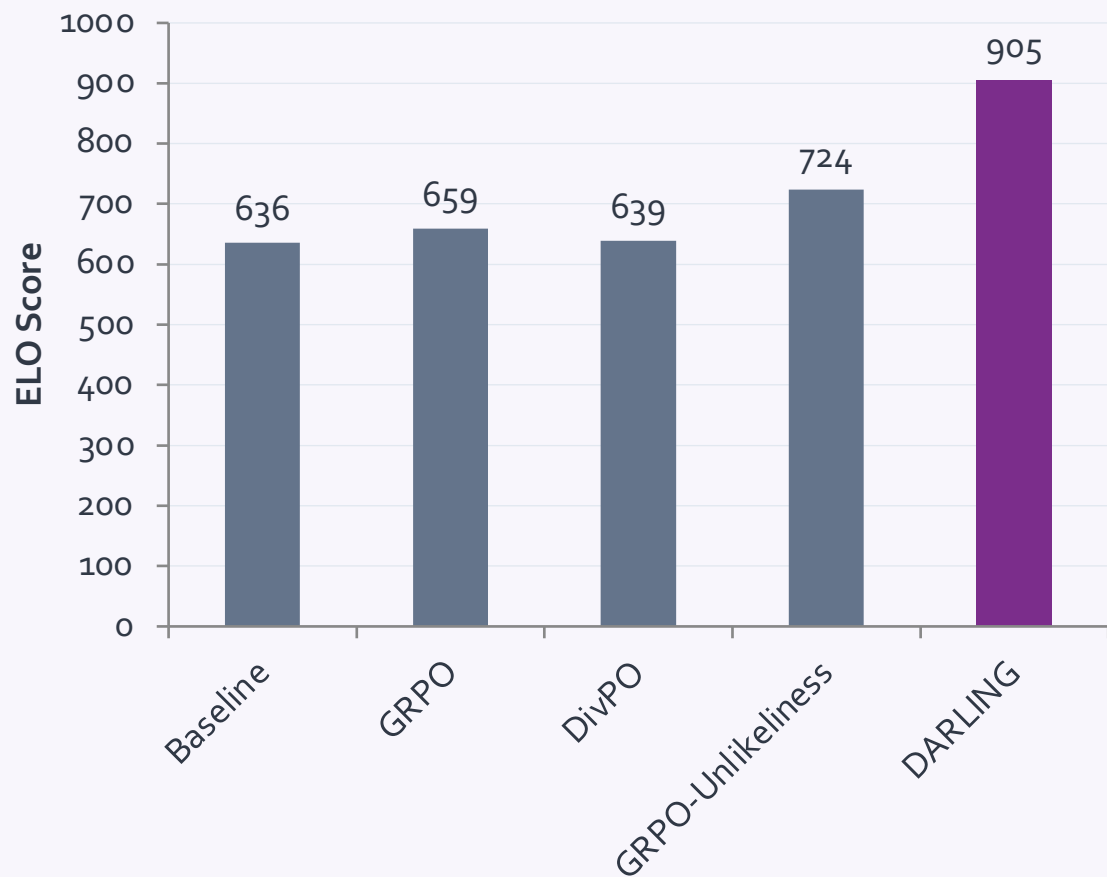
Llama-3.3-70B



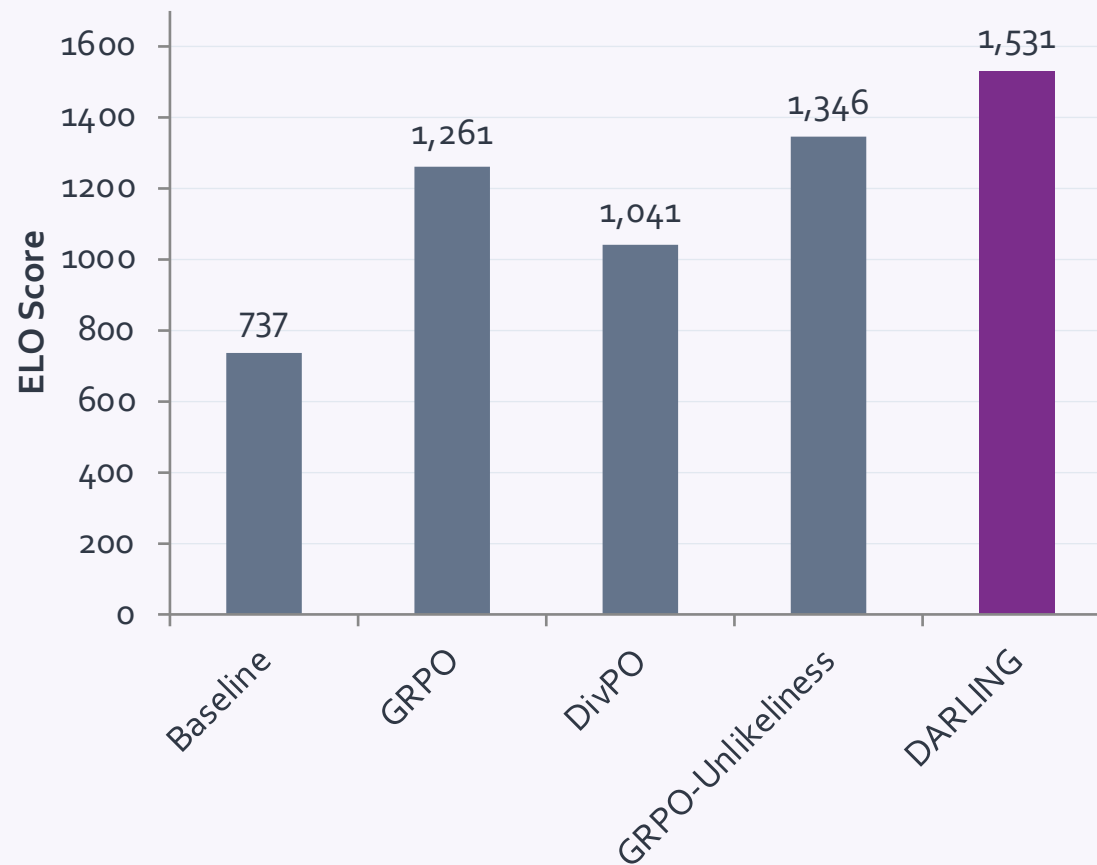
Evaluation on EQ-Bench (Creative writing)

DARLING improves quality and diversity.

Llama-3.1-8B



Llama-3.3-70B



Evaluation on EQ-Bench (Creative writing)

Examples

NoveltyBench [Zhang et al. 2025]; 4 parallel samples; temp=1.0

Input

I am doing online shopping for a new leather wallet. What's the first store/website I should check out? Please just give me one suggestion.

Examples

NoveltyBench [Zhang et al. 2025]; 4 parallel samples; temp=1.0

Input

I am doing online shopping for a new leather wallet. What's the first store/website I should check out? Please just give me one suggestion.

Llama-3.3-70B-Instruct:

Examples

NoveltyBench [Zhang et al. 2025]; 4 parallel samples; temp=1.0

Input

I am doing online shopping for a new leather wallet. What's the first store/website I should check out? Please just give me one suggestion.

Llama-3.3-70B-Instruct:

1. Amazon.

Examples

NoveltyBench [Zhang et al. 2025]; 4 parallel samples; temp=1.0

Input

I am doing online shopping for a new leather wallet. What's the first store/website I should check out? Please just give me one suggestion.

Llama-3.3-70B-Instruct:

1. Amazon.

2. Amazon.

Examples

NoveltyBench [Zhang et al. 2025]; 4 parallel samples; temp=1.0

Input

I am doing online shopping for a new leather wallet. What's the first store/website I should check out? Please just give me one suggestion.

Llama-3.3-70B-Instruct:

1. Amazon.

2. Amazon.

3. Amazon.

Examples

NoveltyBench [Zhang et al. 2025]; 4 parallel samples; temp=1.0

Input

I am doing online shopping for a new leather wallet. What's the first store/website I should check out? Please just give me one suggestion.

Llama-3.3-70B-Instruct:

1. Amazon.

2. Amazon.

3. Amazon.

4. Amazon.

Examples

NoveltyBench [Zhang et al. 2025]; 4 parallel samples; temp=1.0

Input

I am doing online shopping for a new leather wallet. What's the first store/website I should check out? Please just give me one suggestion.

Llama-3.3-70B-Instruct:

1. Amazon.

2. Amazon.

3. Amazon.

4. Amazon.

Llama-3.3-70B-Instruct w/ DARLING:

Examples

NoveltyBench [Zhang et al. 2025]; 4 parallel samples; temp=1.0

Input

I am doing online shopping for a new leather wallet. What's the first store/website I should check out? Please just give me one suggestion.

Llama-3.3-70B-Instruct:

1. Amazon.

2. Amazon.

3. Amazon.

4. Amazon.

Llama-3.3-70B-Instruct w/ DARLING:

1. Fossil. They specialize in leather goods, offer a huge variety (slim to traditional, many ...

2. Bellroy. They're highly regarded not just for quality leather that ages well, but also very ...

3. Coach.com. They're not just a brand name; consistently, their site offers high-quality

4. Bellroy. They specialize in leather goods, offer a wide, well-curated range of wallet ...

Summary

- **Motivation:** Standard post-training “sharpens” outputs
 - (i.e., collapses semantic diversity).
 - This makes models less useful for creative, exploratory tasks.
- **DARLING:** an *online* RL framework that multiplies a quality reward by a learned *semantic* diversity score.
- **Finding:** A Pareto improvement (higher quality and higher diversity) across benchmarks.

Summary

- **Motivation:** Standard post-training “sharpens” outputs
 - (i.e., collapses semantic diversity).
 - This makes models less useful for creative, exploratory tasks.
- **DARLING:** an *online* RL framework that multiplies a quality reward by a learned *semantic* diversity score.
- **Finding:** A Pareto improvement (higher quality and higher diversity) across benchmarks.

Summary

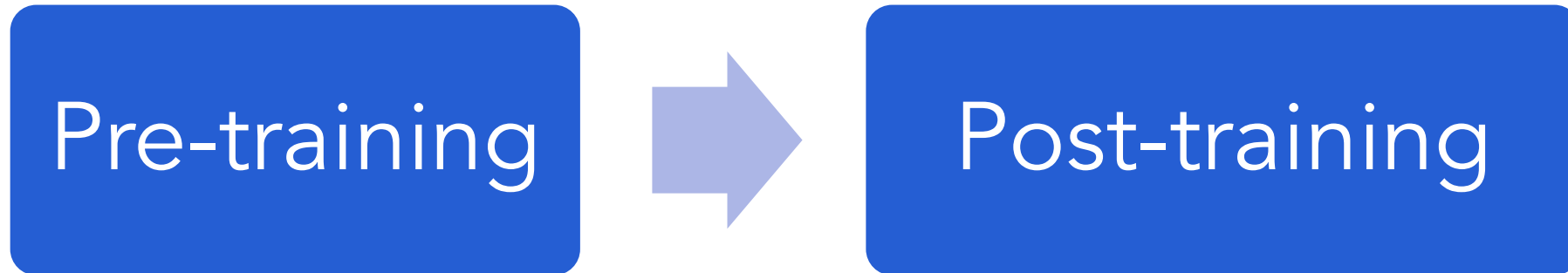
- **Motivation:** Standard post-training “sharpens” outputs
 - (i.e., collapses semantic diversity).
 - This makes models less useful for creative, exploratory tasks.
- **DARLING:** an *online* RL framework that multiplies a quality reward by a learned *semantic* diversity score.
- **Finding:** A Pareto improvement (higher quality and higher diversity) across benchmarks.

Summary

- **Motivation:** Standard post-training “sharpens” outputs
 - (i.e., collapses semantic diversity).
 - This makes models less useful for creative, exploratory tasks.
- **DARLING:** an *online* RL framework that multiplies a quality reward by a learned *semantic* diversity score.
- **Finding:** A Pareto improvement (higher quality and higher diversity) across benchmarks.

Closing the Loop: Rethinking the Reasoning Stack

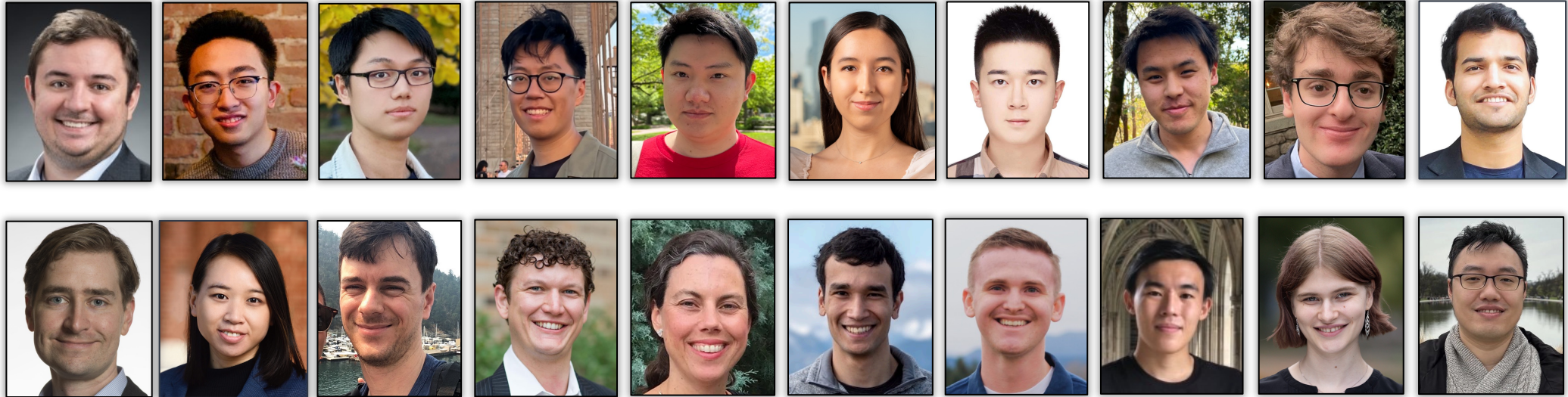
- The success of this paradigm has been remarkable.
- We have a remarkable castle. Now it's time to rethink the bricks.



Closing the Loop: Rethinking the Reasoning Stack

- A few speculations/questions:
 - Pre-training [continue to] do the heavy-lifting. We need more research here!
 - The separation between pre-training and post-training may be artificial.
 - There is a lot of non-lang out there. Nature is full of “intelligence”.
How can we bring them together?

Thanks to our wonderful collaborators!



Sponsors:

