MetaICL: Learning to Learn In Context

Sewon Min, Mike Lewis, Luke Zettlemoyer, Hannaneh Hajishirzi
University of Washington, Meta AI, Allen Institute for AI




UW NLP Meta AI2 NAACL 2022

In-Context Learning

In-Context Learning

***k* shot data**

Input: Circulation revenue has increased by 5% in Finland. Output : Positive

Input: Panostaja did not disclose the purchase price. Output: Neutral

Input: Paying off the national debt will be extremely painful. Output: Negative

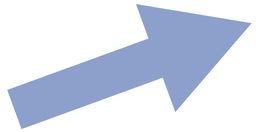
In-Context Learning

***k* shot data**

Input: Circulation revenue has increased by 5% in Finland. Output : Positive

Input: Panostaja did not disclose the purchase price. Output: Neutral

Input: Paying off the national debt will be extremely painful. Output: Negative



Concatenate

$x =$ Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

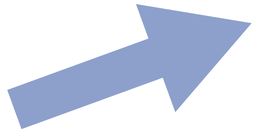
In-Context Learning

***k* shot data**

Input: Circulation revenue has increased by 5% in Finland. Output : Positive

Input: Panostaja did not disclose the purchase price. Output: Neutral

Input: Paying off the national debt will be extremely painful. Output: Negative



Concatenate

$x =$ Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. //

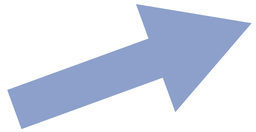
In-Context Learning

***k* shot data**

Input: Circulation revenue has increased by 5% in Finland. Output : Positive

Input: Panostaja did not disclose the purchase price. Output: Neutral

Input: Paying off the national debt will be extremely painful. Output: Negative



Concatenate

$x =$ Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. //

Use LM

Language Model

In-Context Learning

***k* shot data**

Input: Circulation revenue has increased by 5% in Finland. Output : Positive

Input: Panostaja did not disclose the purchase price. Output: Neutral

Input: Paying off the national debt will be extremely painful. Output: Negative

Concatenate

$x =$ Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. //

Use LM

Language Model

Scoring

$P_{\text{LM}}(\text{“positive”}|x) > P_{\text{LM}}(\text{“negative”}|x)$

In-Context Learning

***k* shot data**

Input: Circulation revenue has increased by 5% in Finland. Output : Positive

Input: Panostaja did not disclose the purchase price. Output: Neutral

Input: Paying off the national debt will be extremely painful. Output: Negative

Concatenate

$x =$ Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. //

Use LM

Language Model

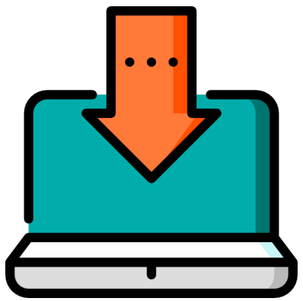
Scoring

$P_{\text{LM}}(\text{“positive”}|x) > P_{\text{LM}}(\text{“negative”}|x)$

In-context learning
(Brown et al. 2020)

GPT-3

In-context learning
(Brown et al. 2020)

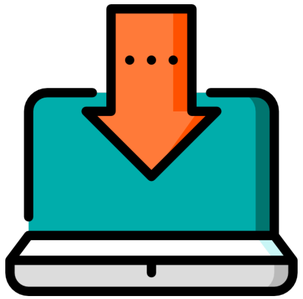


Better tuning of prompt
(Shin et al. 2020, Jiang et al. 2020, Gao et al. 2021)

GPT-3

In-context learning
(Brown et al. 2020)

Better scoring
(Zhao et al. 2021, Holtzman et al. 2021)

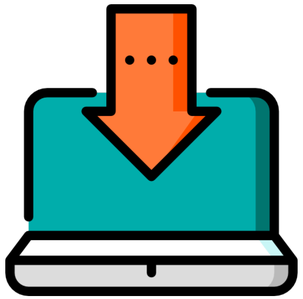


Better tuning of prompt
(Shin et al. 2020, Jiang et al. 2020, Gao et al. 2021)

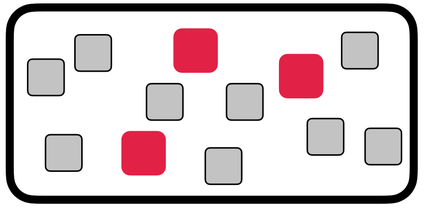
GPT-3

In-context learning
(Brown et al. 2020)

Better scoring
(Zhao et al. 2021, Holtzman et al. 2021)



Better tuning of prompt
(Shin et al. 2020, Jiang et al. 2020, Gao et al. 2021)



Better choice of in-context examples
(Liu et al. 2021, Lu et al. 2021, Rubin et al. 2021)

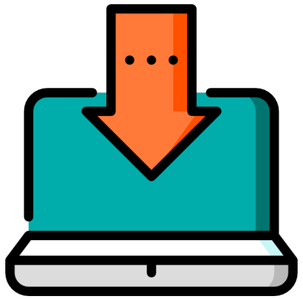
Next presentation!

We always used the plain language model

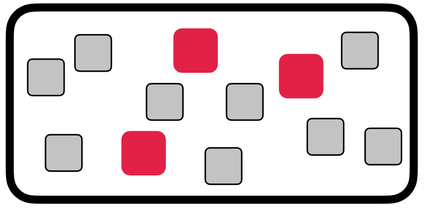
GPT-3

In-context learning
(Brown et al. 2020)

Better scoring
(Zhao et al. 2021, Holtzman et al. 2021)



Better tuning of prompt
(Shin et al. 2020, Jiang et al. 2020, Gao et al. 2021)



Better choice of in-context examples
(Liu et al. 2021, Lu et al. 2021, Rubin et al. 2021)

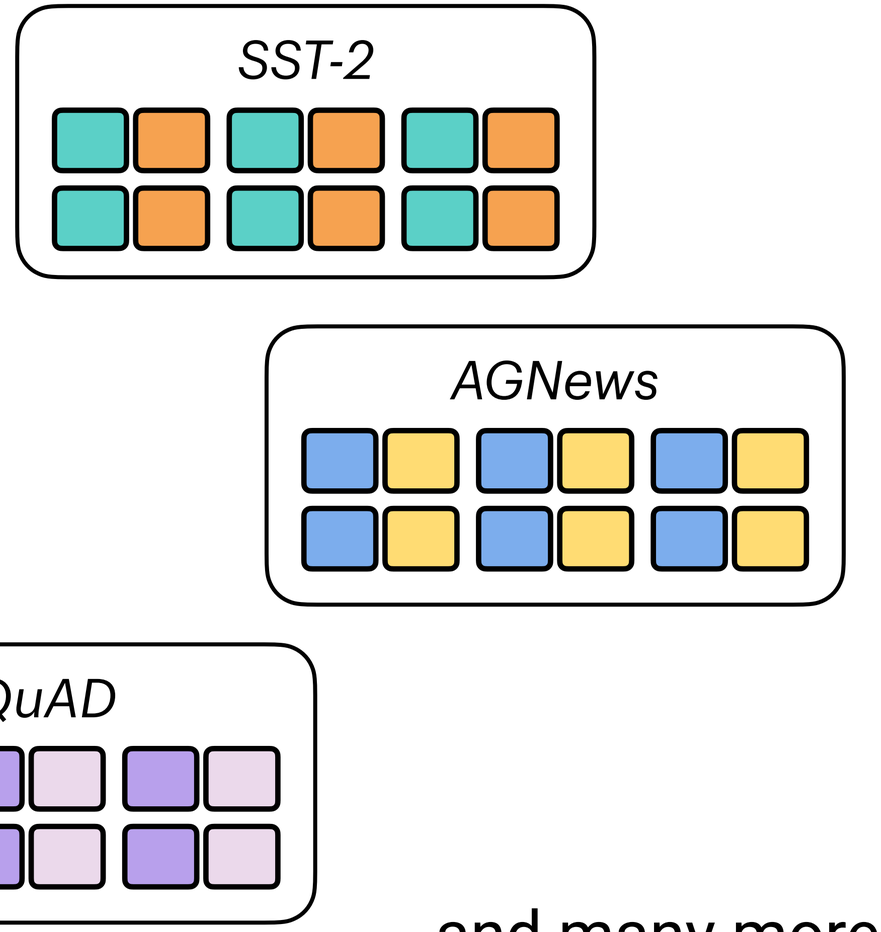
Next presentation!

We always used the plain language model

*“Can we **meta-train** the model to be **a better in-context learner**?”*

We always used the plain language model

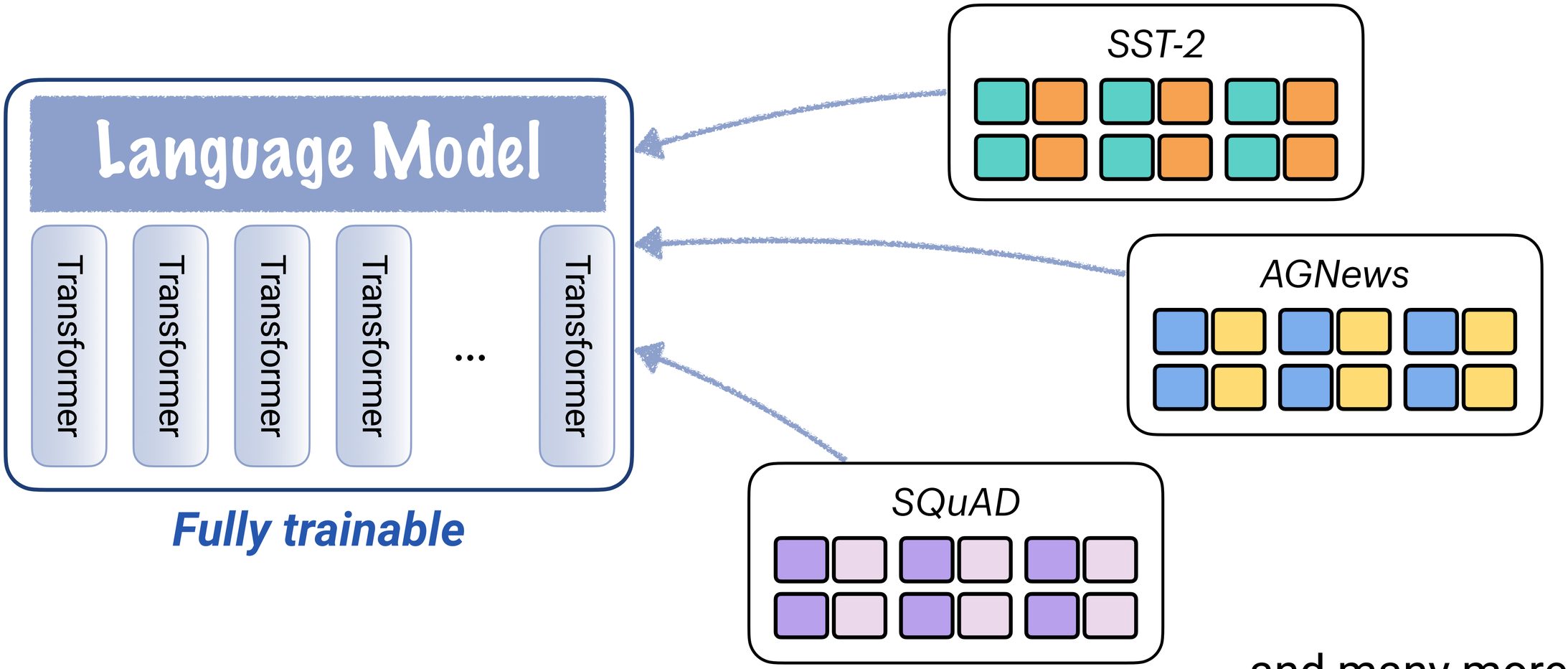
*"Can we **meta-train** the model to be **a better in-context learner**?"*



... and many more

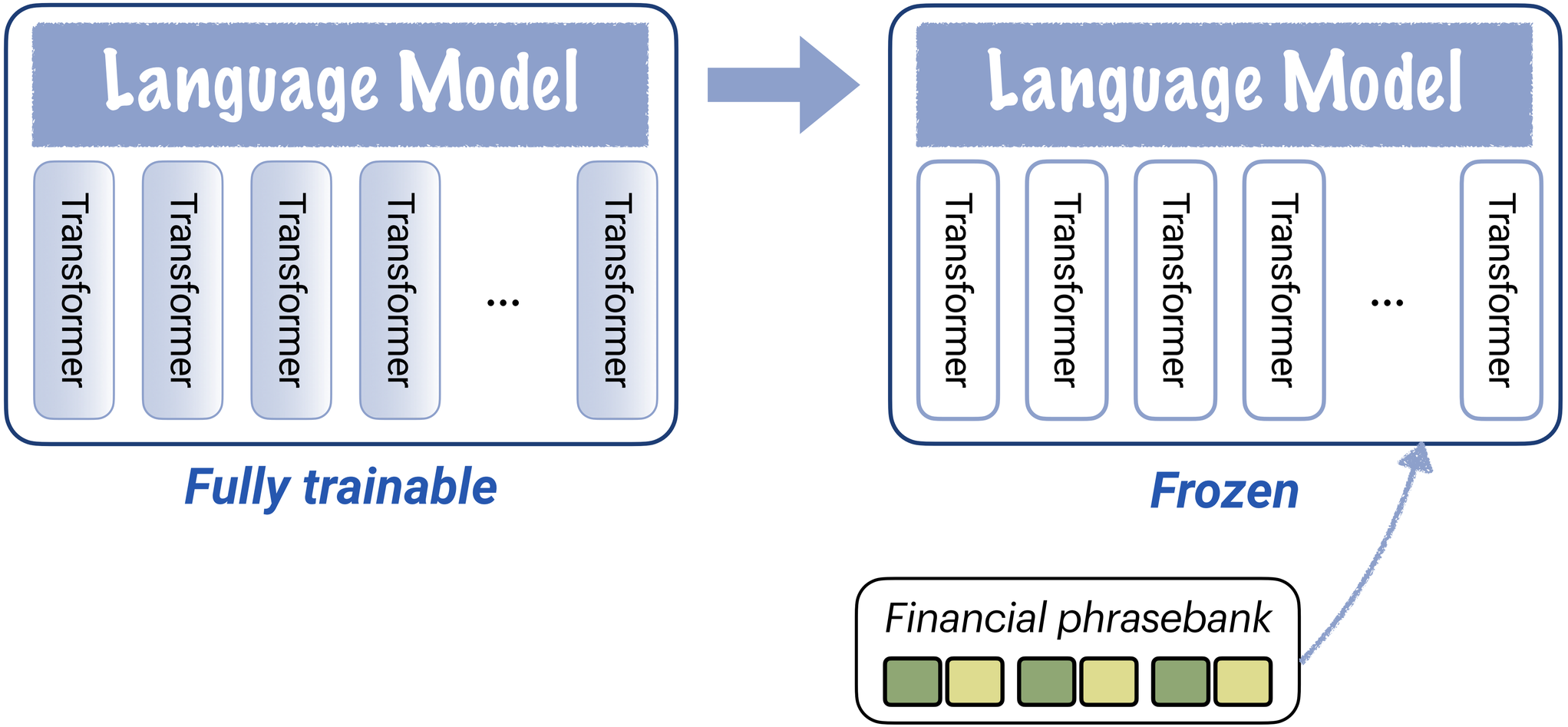
We always used the plain language model

*"Can we **meta-train** the model to be **a better in-context learner**?"*



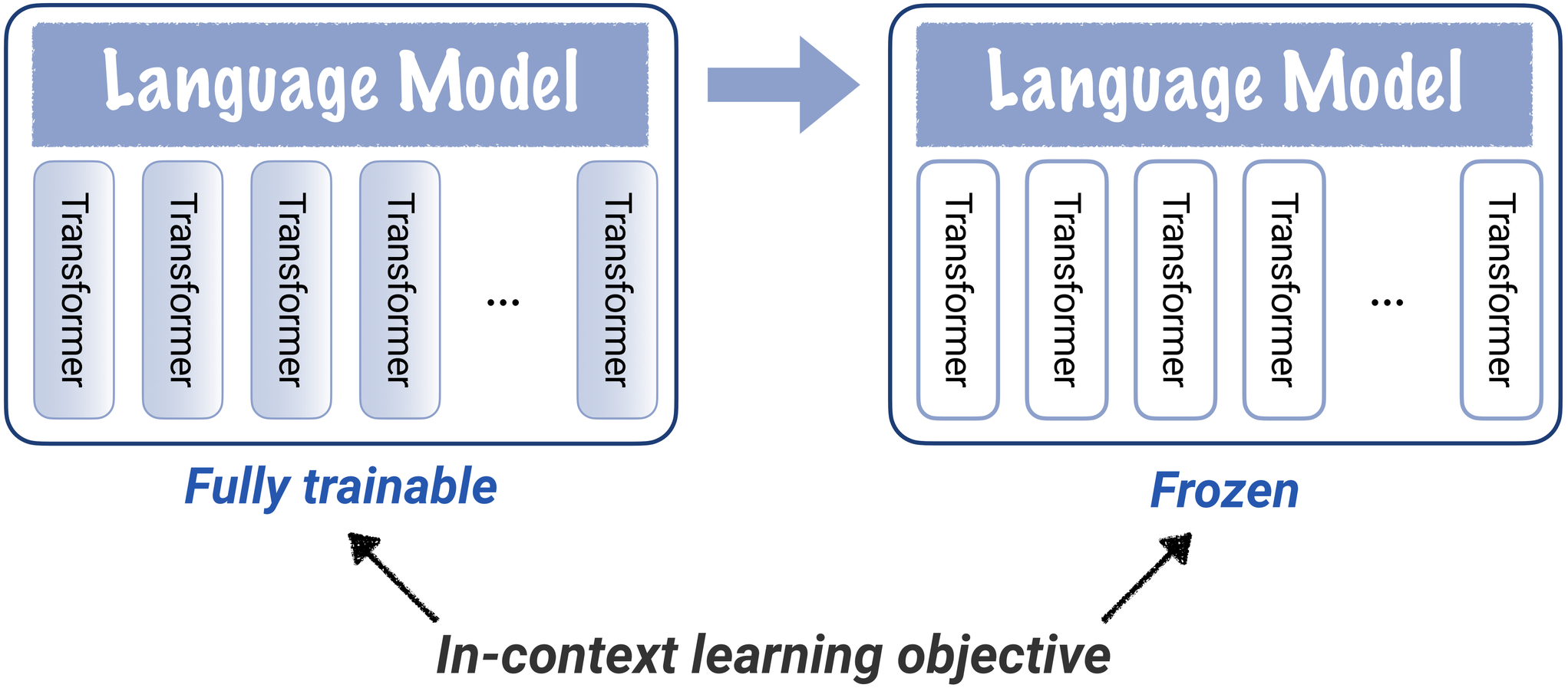
"We always used the plain language model"

*"Can we **meta-train** the model to be **a better in-context learner**?"*



“We always used the plain language model”

*“Can we **meta-train** the model to be **a better in-context learner**?”*



MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$

MetaICL

A collection of meta-training tasks

SST-2 *AGNew* *SQuAD* ...

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$

MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

A collection of meta-training tasks

SST-2 | *AGNew* | *SQuAD* | ...

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$

MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$

A collection of meta-training tasks

SST-2 | *AGNew* | *SQuAD* | ⋯

(x_1, y_1) (x_2, y_2) ⋯ (x_k, y_k) (x_{k+1}, y_{k+1})

MetaICL

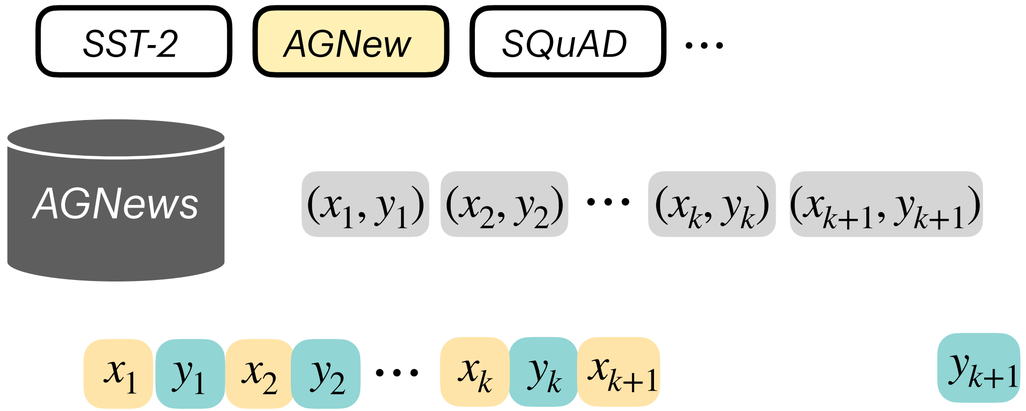
Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$

A collection of meta-training tasks



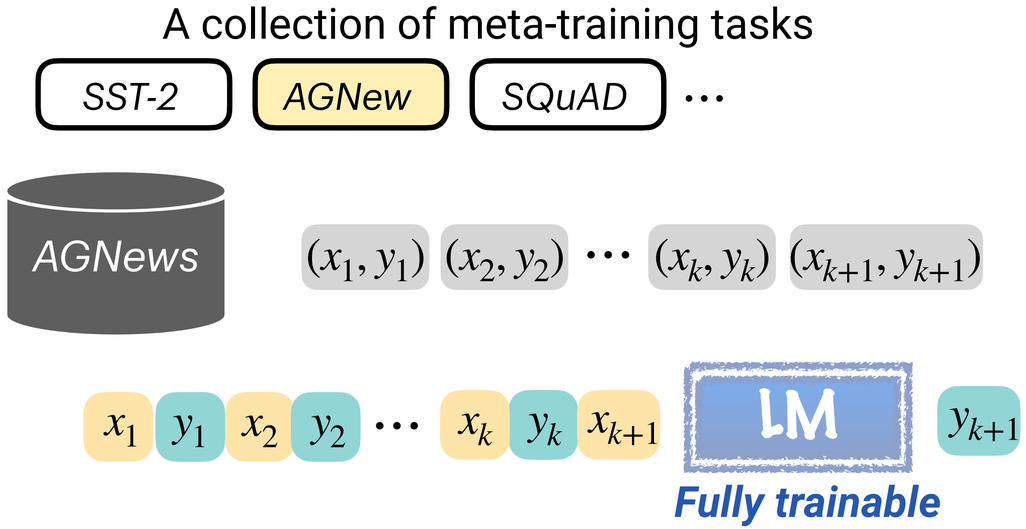
MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$



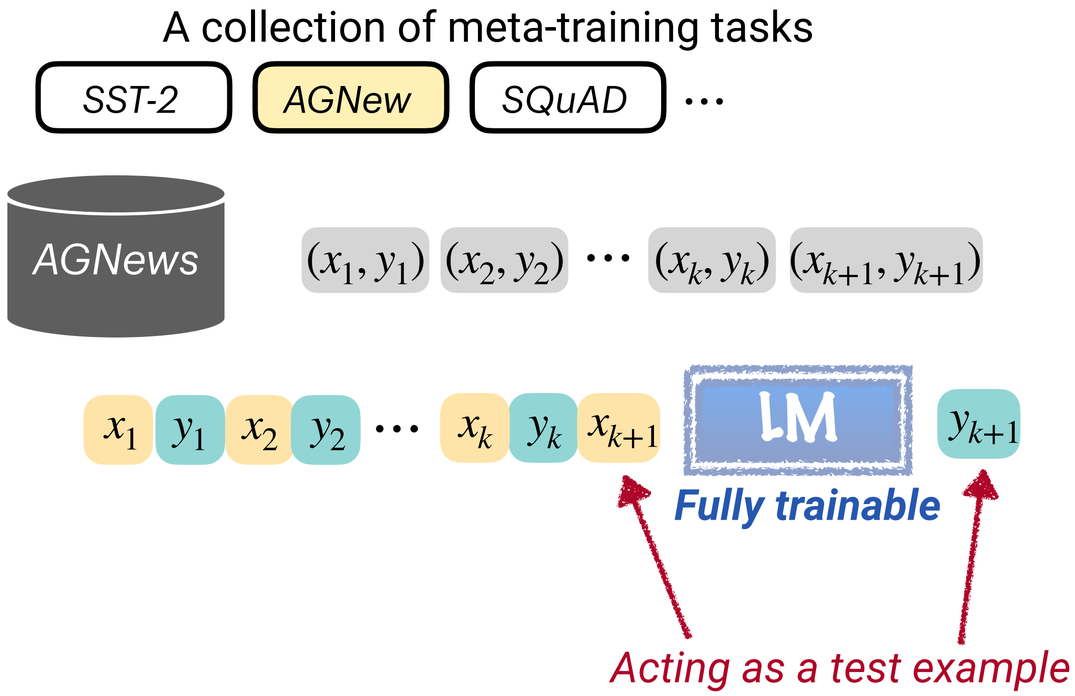
MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$



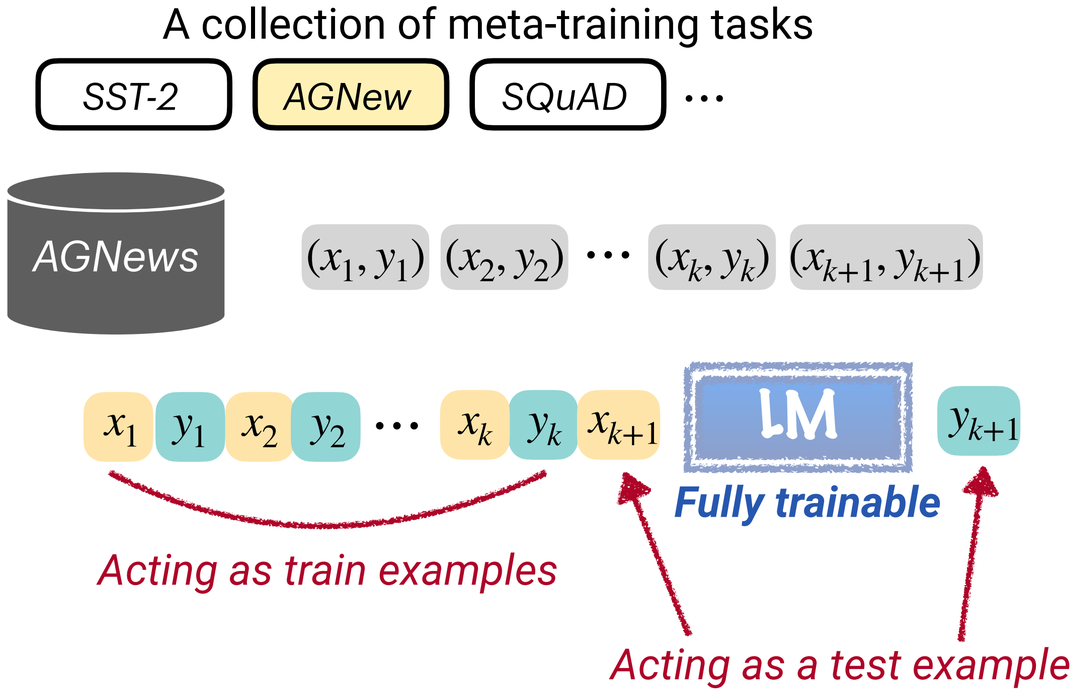
MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$



MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$

A collection of meta-training tasks

SST-2 *AGNew* *SQuAD* ⋯

(x_1, y_1) (x_2, y_2) ⋯ (x_k, y_k) (x_{k+1}, y_{k+1})

x_1 y_1 x_2 y_2 ⋯ x_k y_k x_{k+1} **LM** y_{k+1}

Fully trainable

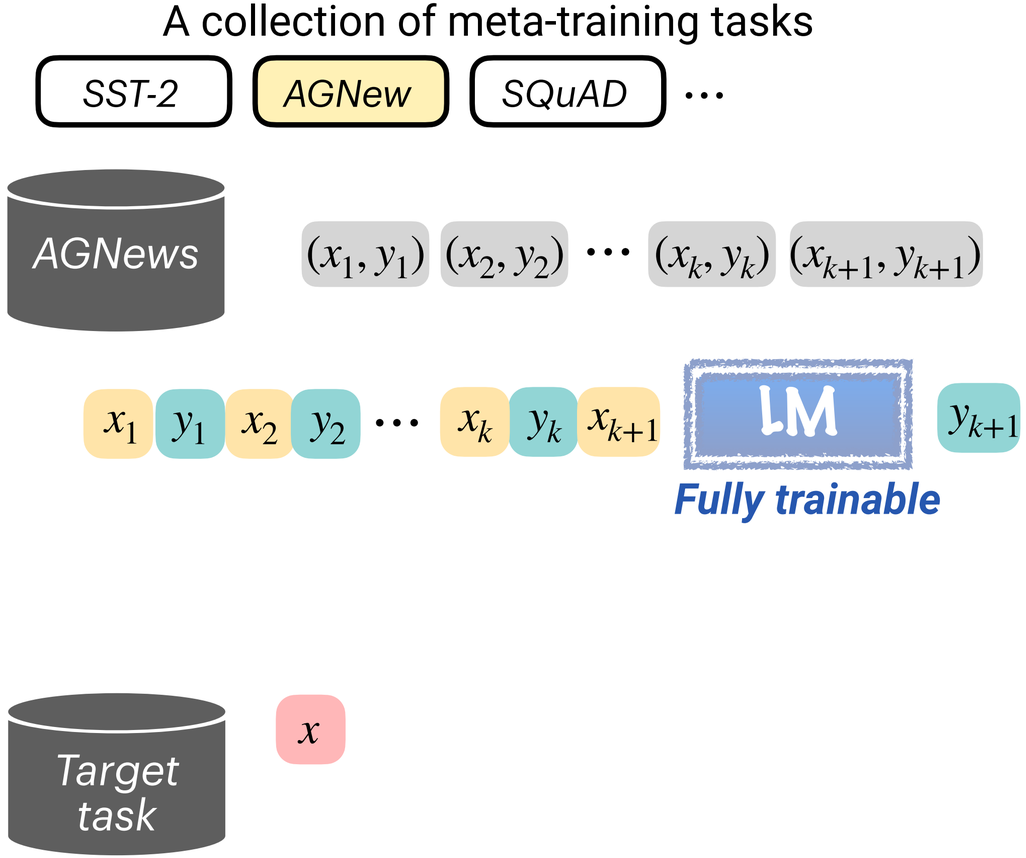
MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$



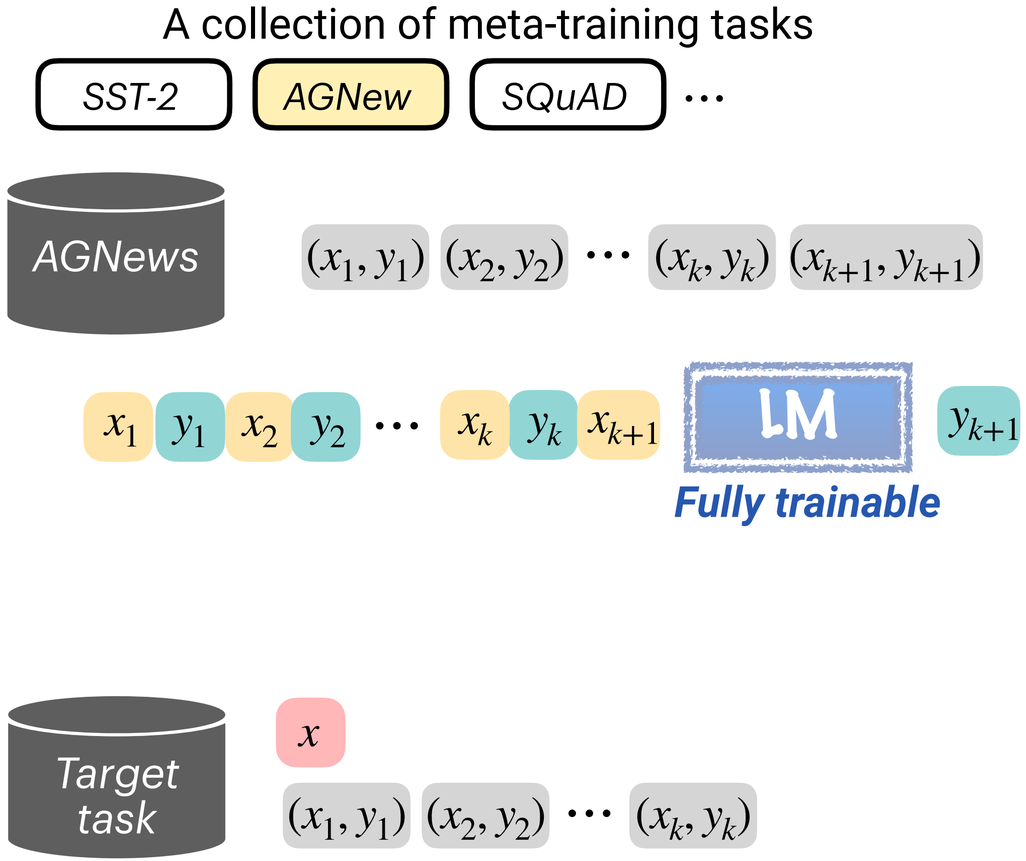
MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$



MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$

A collection of meta-training tasks

SST-2 | AGNew | SQuAD | ⋯

(x_1, y_1) (x_2, y_2) ⋯ (x_k, y_k) (x_{k+1}, y_{k+1})

x_1 y_1 x_2 y_2 ⋯ x_k y_k x_{k+1} **LM** y_{k+1}

Fully trainable

x

(x_1, y_1) (x_2, y_2) ⋯ (x_k, y_k)

x_1 y_1 x_2 y_2 ⋯ x_k y_k x

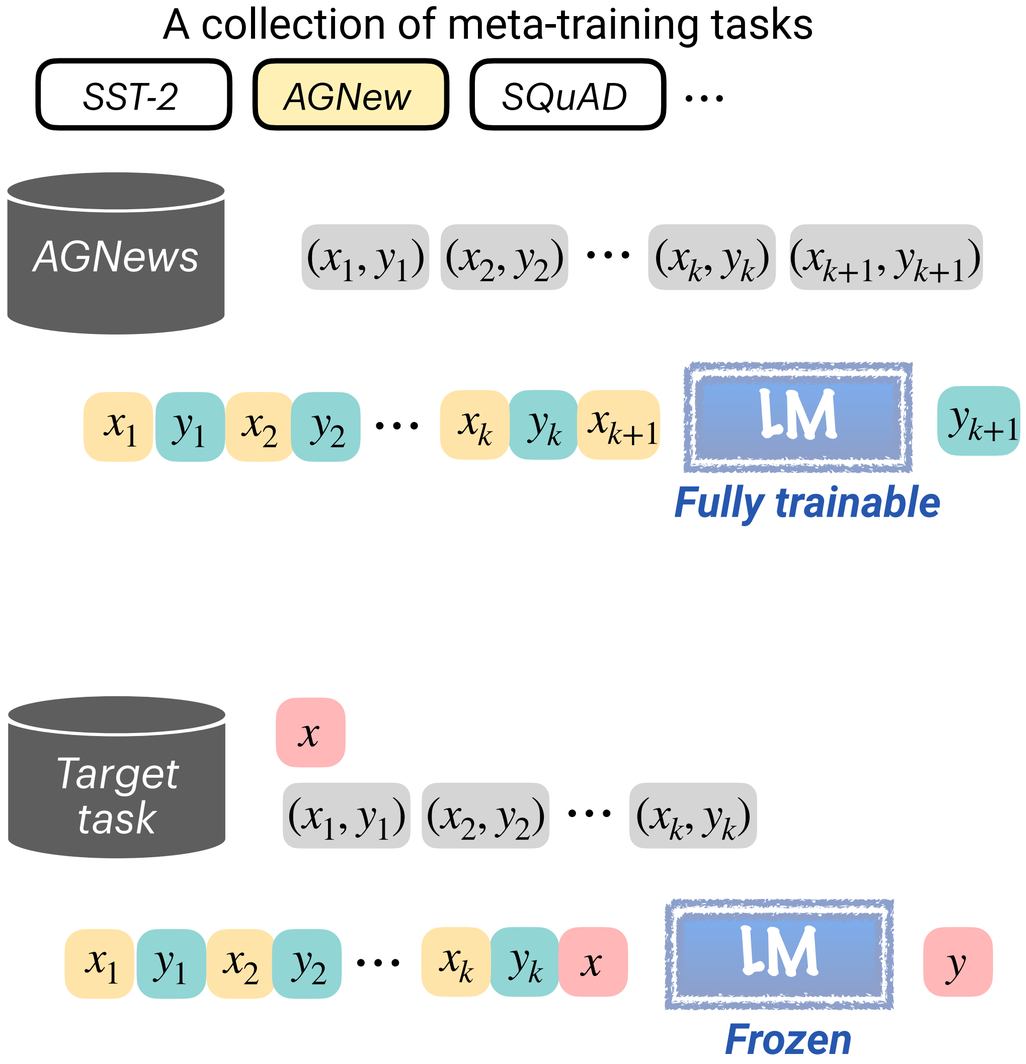
MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$



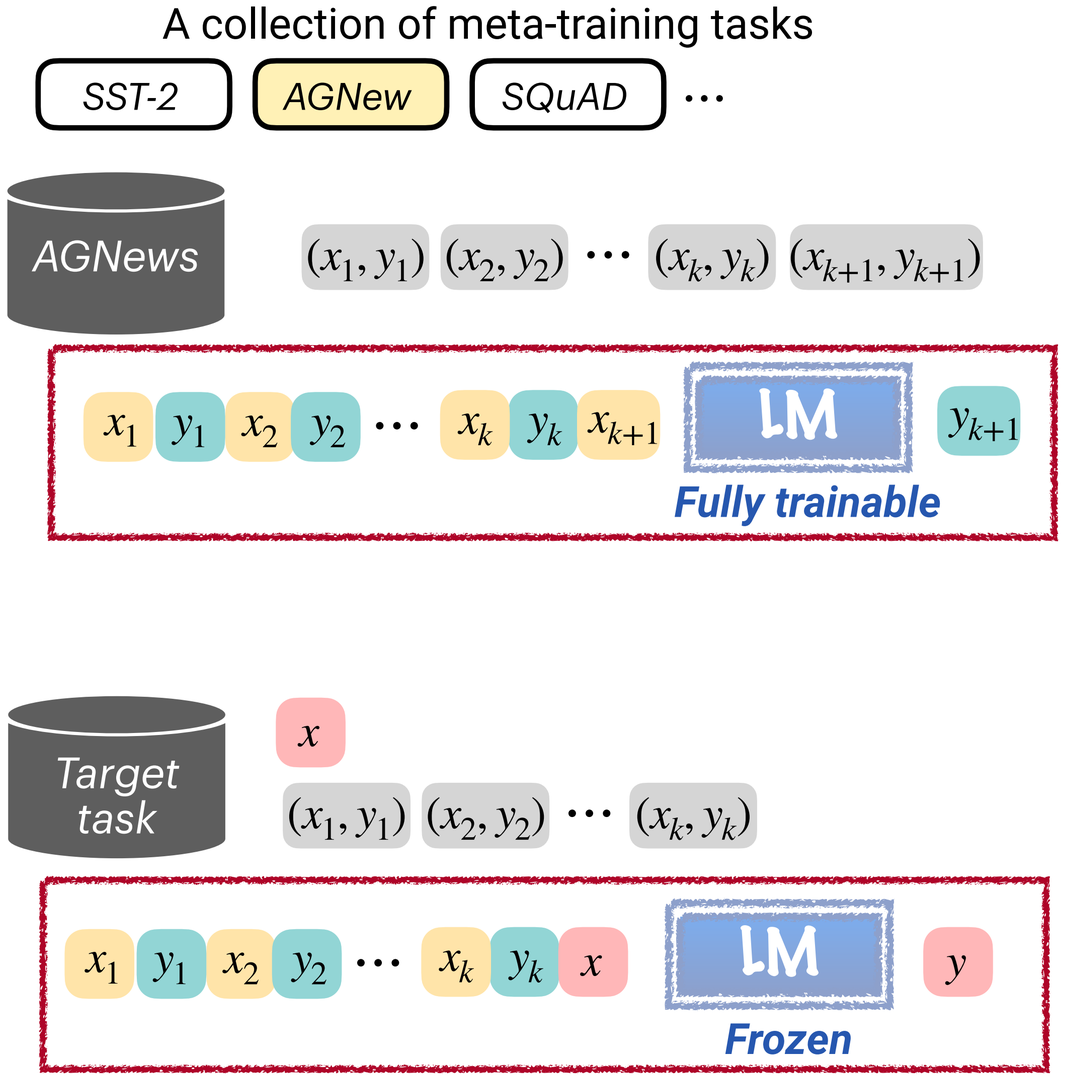
MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$



MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$

Direct MetaICL

MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(y_{k+1} \mid x_1, y_1, \cdots, x_k, y_k, x_{k+1})$

Inference

Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$

Direct MetaICL

Meta-training

1. Sample a task T
2. Sample $(x_1, y_1), \cdots, (x_{k+1}, y_{k+1})$ from T
3. Maximize $\log P(x_{k+1} \mid y_1, x_1, \cdots, y_k, x_k, y_{k+1})$

Inference

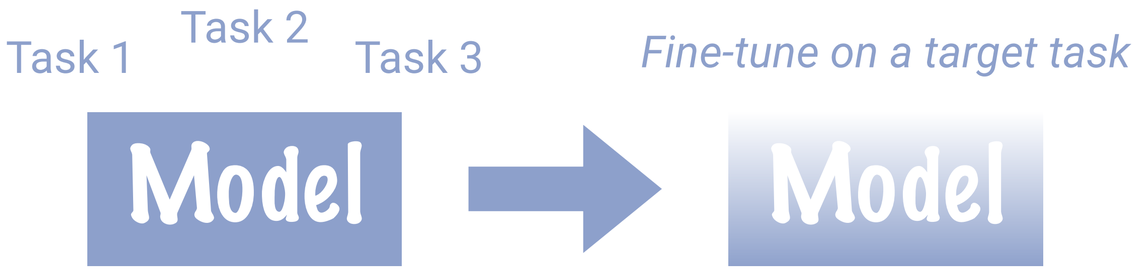
Given training examples $(x_1, y_1), \cdots, (x_k, y_k)$,
find $\text{argmax}_{y \in \mathcal{Y}} P(x \mid y_1, x_1, \cdots, y_k, x_k, y)$

Channel MetaICL

"Noisy Channel Language Model Prompting for Few-Shot Text Classification" ACL 2022

Related Work

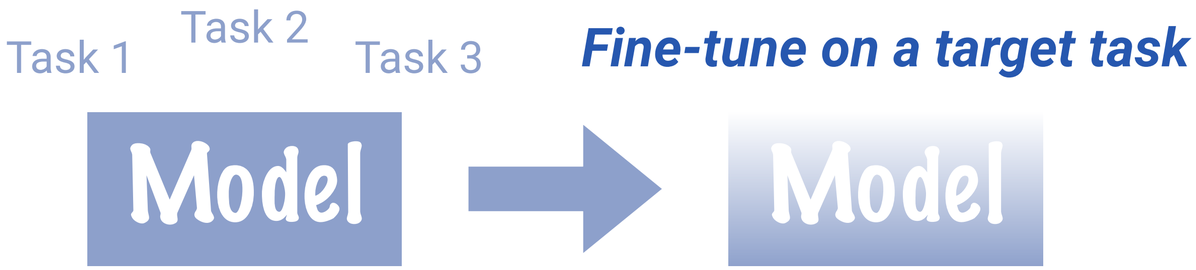
Related Work



Multi-task learning/Meta-learning

(Villa and Drissi 2002, Evgeniou and Pontil 2004, Ruder 2017, Wang et al. 2020, Aghajanyan et al. 2021)

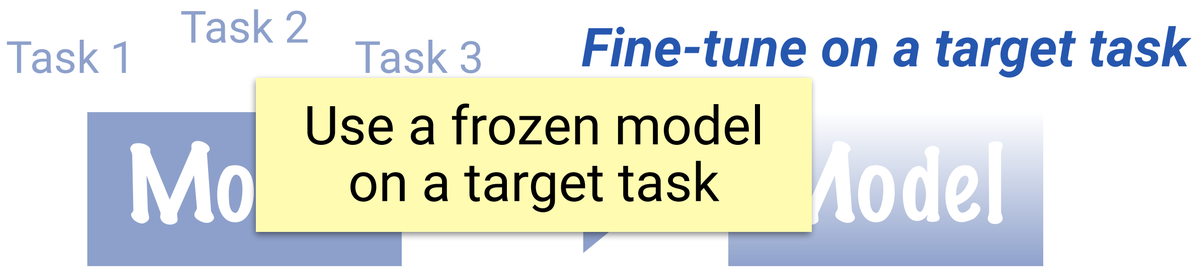
Related Work



Multi-task learning/Meta-learning

(Villa and Drissi 2002, Evgeniou and Pontil 2004, Ruder 2017, Wang et al. 2020, Aghajanyan et al. 2021)

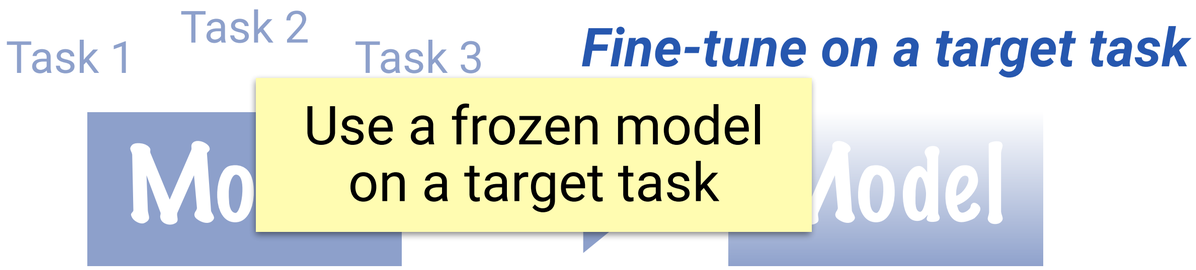
Related Work



Multi-task learning/Meta-learning

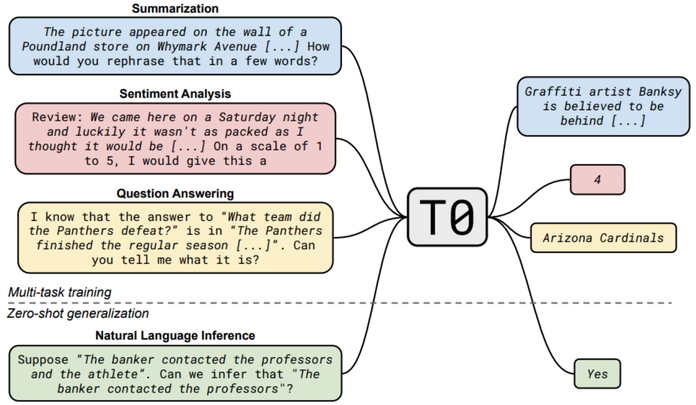
(Villa and Drissi 2002, Evgeniou and Pontil 2004, Ruder 2017, Wang et al. 2020, Aghajanyan et al. 2021)

Related Work



Multi-task learning/Meta-learning

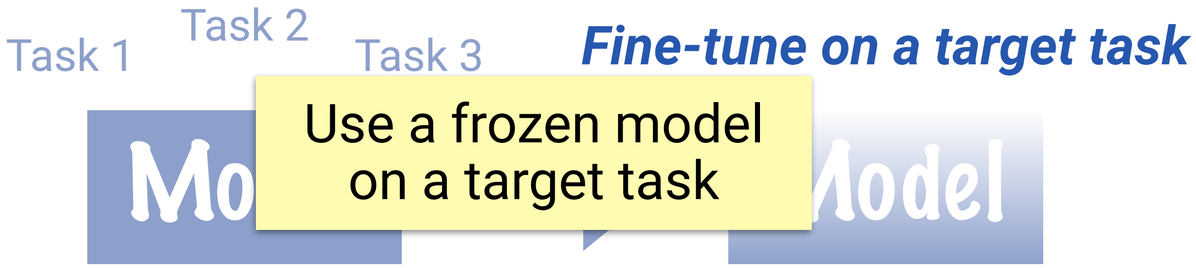
(Villa and Drissi 2002, Evgeniou and Pontil 2004, Ruder 2017, Wang et al. 2020, Aghajanyan et al. 2021)



Instruction-reading models

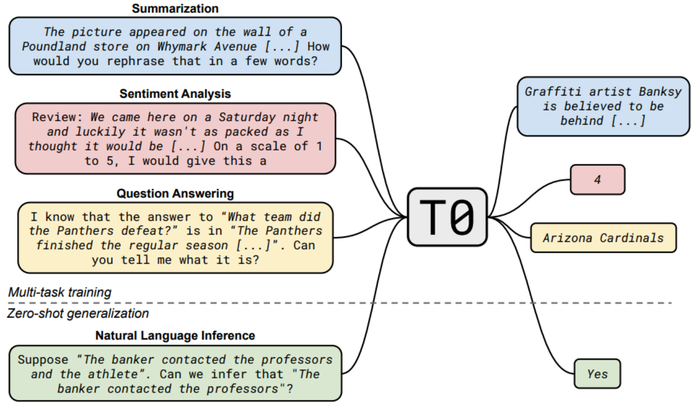
(Mishra et al. 2021, Wei et al. 2021, Sanh et al. 2021, Zhong et al. 2021)
Figure from Sanh et al. 2021

Related Work



Multi-task learning/Meta-learning

(Villa and Drissi 2002, Evgeniou and Pontil 2004, Ruder 2017, Wang et al. 2020, Aghajanyan et al. 2021)

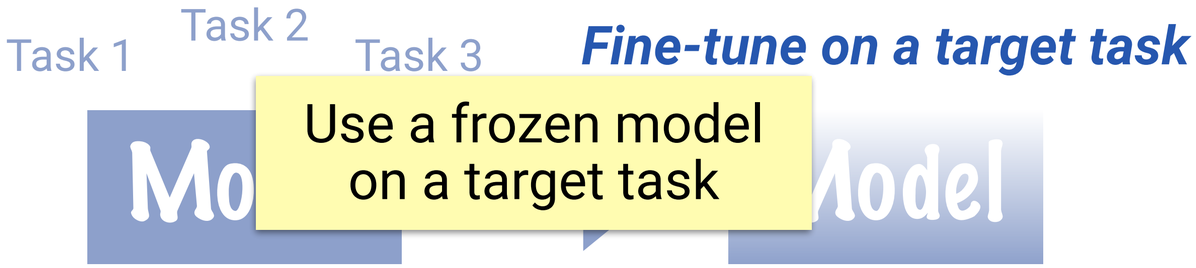


Instruction-reading models

(Mishra et al. 2021, Wei et al. 2021, Sanh et al. 2021, Zhong et al. 2021)
Figure from Sanh et al. 2021

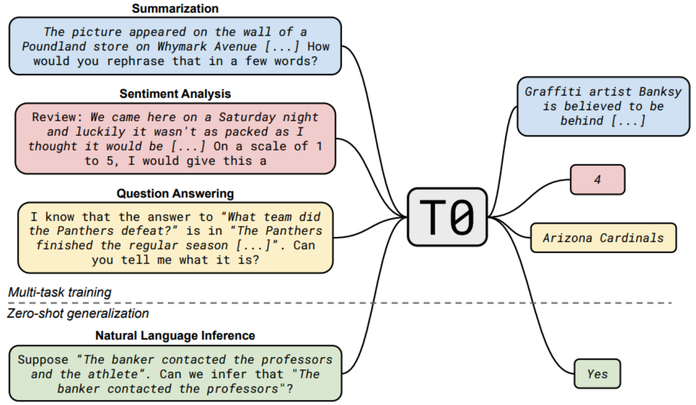
Define the task as the input

Related Work



Multi-task learning/Meta-learning

(Villa and Drissi 2002, Evgeniou and Pontil 2004, Ruder 2017, Wang et al. 2020, Aghajanyan et al. 2021)



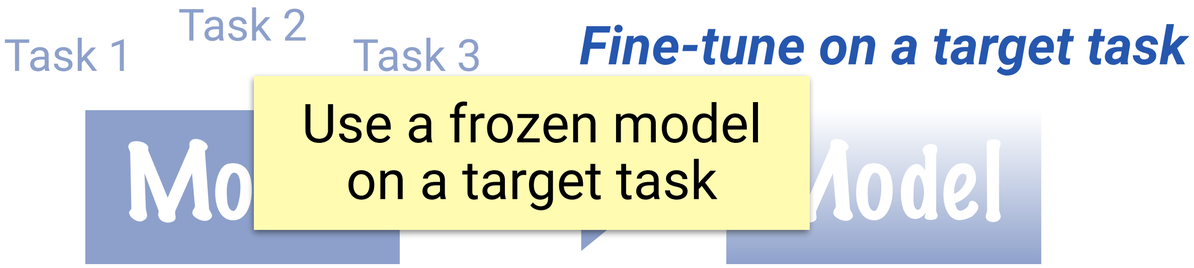
Instruction-reading models

(Mishra et al. 2021, Wei et al. 2021, Sanh et al. 2021, Zhong et al. 2021)
Figure from Sanh et al. 2021

Define the task as the input

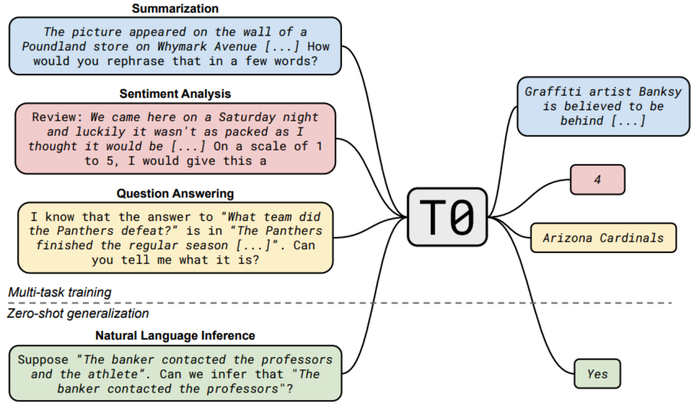
Natural language description vs. *k*-shot data

Related Work



Multi-task learning/Meta-learning

(Villa and Drissi 2002, Evgeniou and Pontil 2004, Ruder 2017, Wang et al. 2020, Aghajanyan et al. 2021)



Instruction-reading models

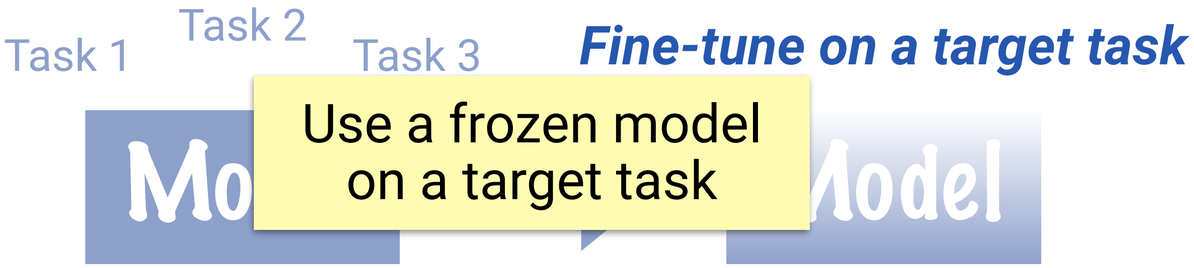
(Mishra et al. 2021, Wei et al. 2021, Sanh et al. 2021, Zhong et al. 2021)
Figure from Sanh et al. 2021

Define the task as the input

Natural language description vs. *k*-shot data

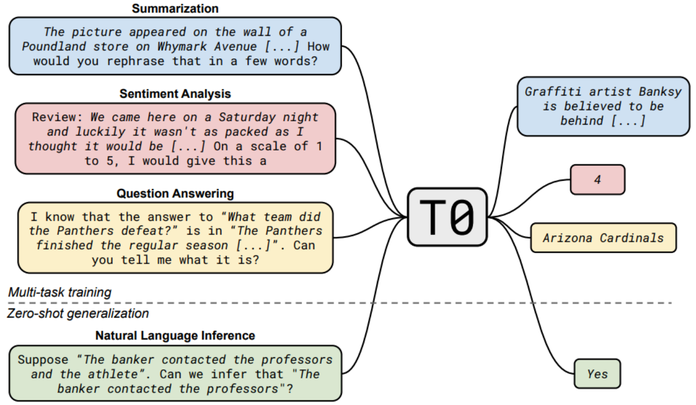
Empirically, MetaICL can be better/complementary

Related Work



Multi-task learning/Meta-learning

(Villa and Drissi 2002, Evgeniou and Pontil 2004, Ruder 2017, Wang et al. 2020, Aghajanyan et al. 2021)



Instruction-reading models

(Mishra et al. 2021, Wei et al. 2021, Sanh et al. 2021, Zhong et al. 2021)
Figure from Sanh et al. 2021

Instruction: "*Is the comment positive?*"
x1: "*Good movie!*" y1: "*yes*"
x2: "*Bad movie!*" y2: "*no*"

$\theta := \theta - \Delta$

Concurrent work

Chen et al. 2021
"Meta-learning via Language Model In-context Tuning"

Define the task as the input

Natural language description vs. *k*-shot data

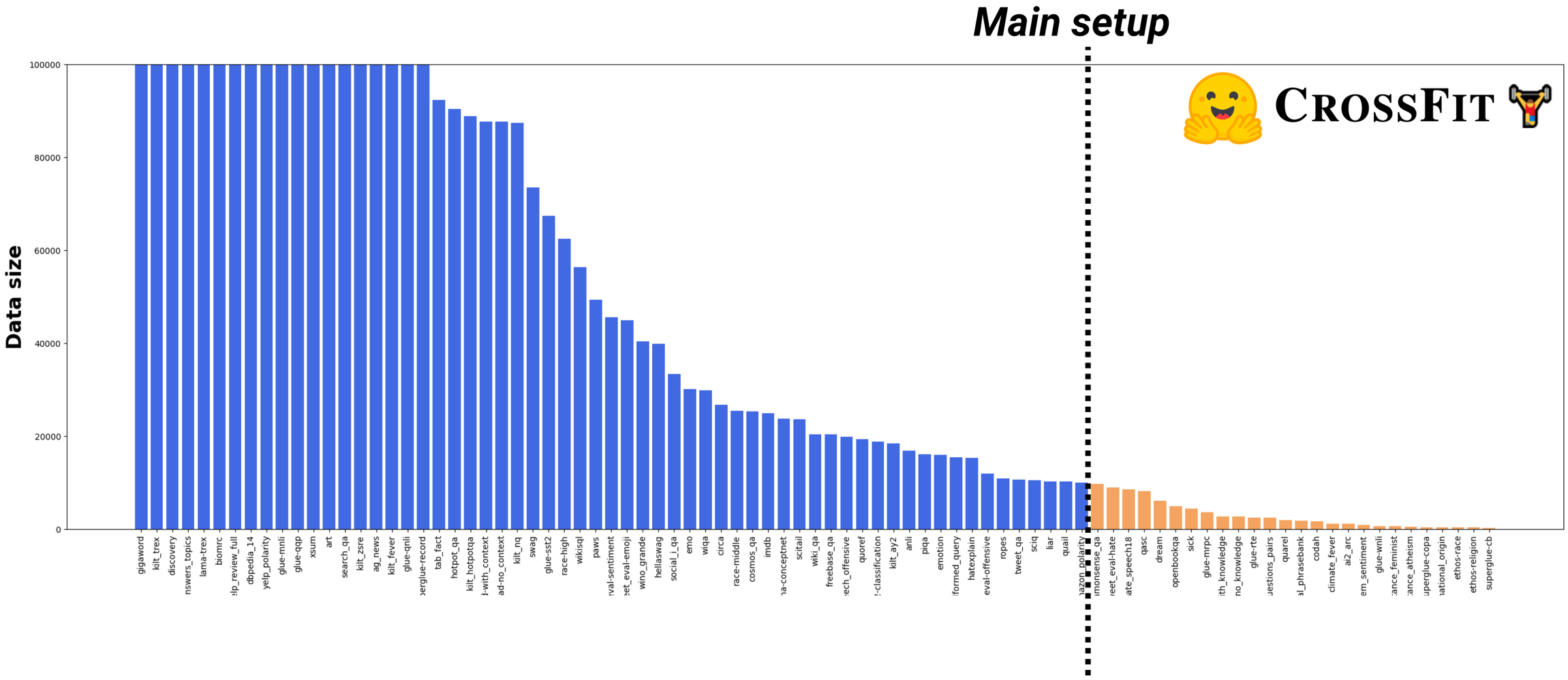
Empirically, MetaICL can be better/complementary

Experimental Setup

In total, experiment with **142** NLP datasets and **7** different metatrain-target splits

Experimental Setup

In total, experiment with **142** NLP datasets and **7** different metatrain-target splits

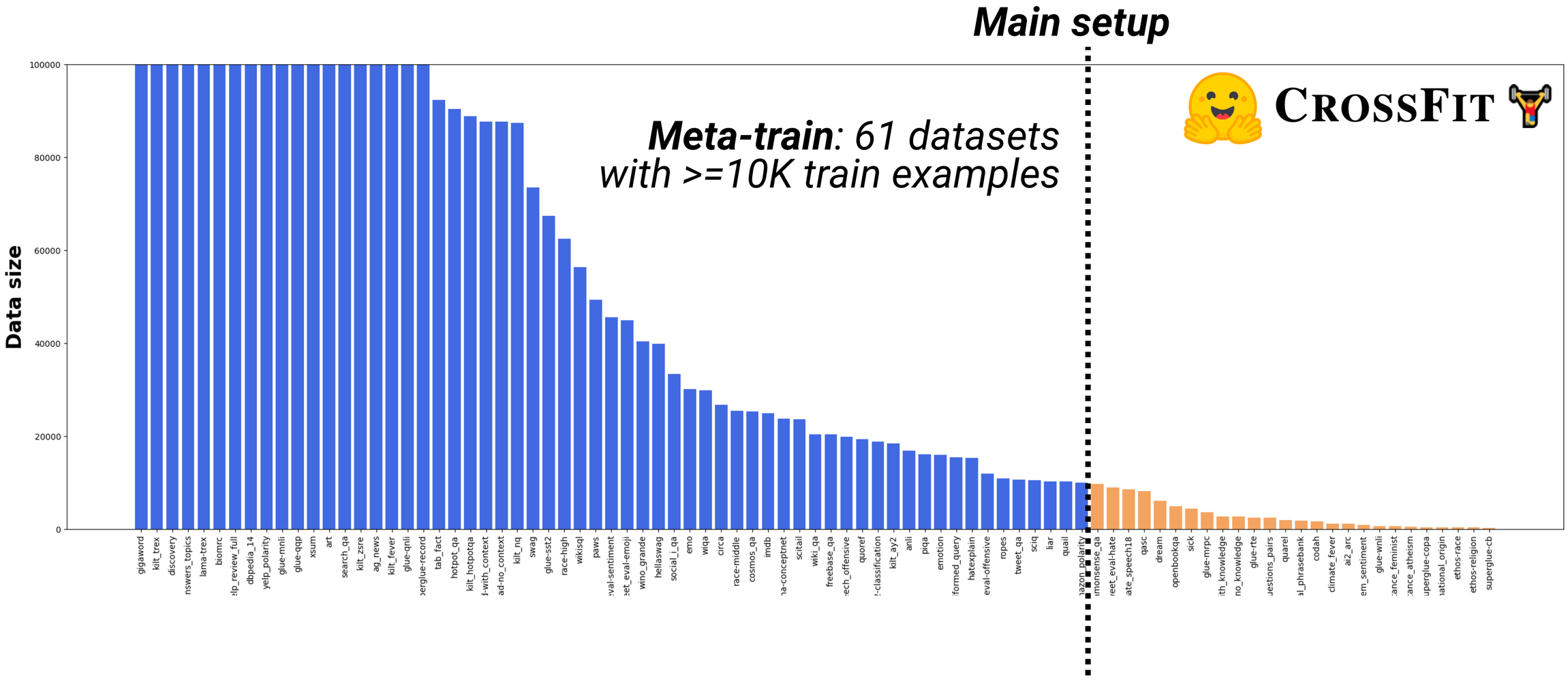


See the paper for more setups!

(Lhoest et al. 2021, Ye et al. 2021)

Experimental Setup

In total, experiment with **142** NLP datasets and **7** different metatrain-target splits

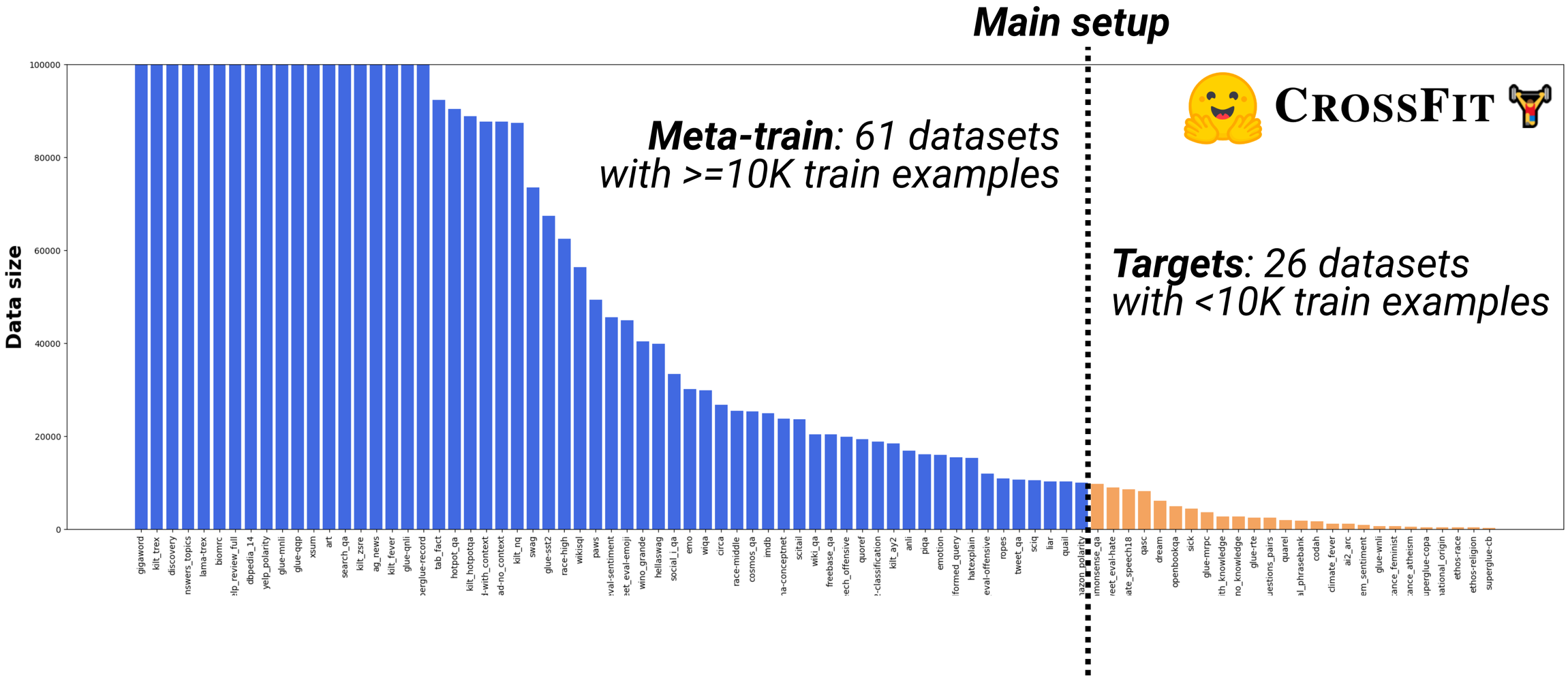


See the paper for more setups!

(Lhoest et al. 2021, Ye et al. 2021)

Experimental Setup

In total, experiment with **142** NLP datasets and **7** different metatrain-target splits

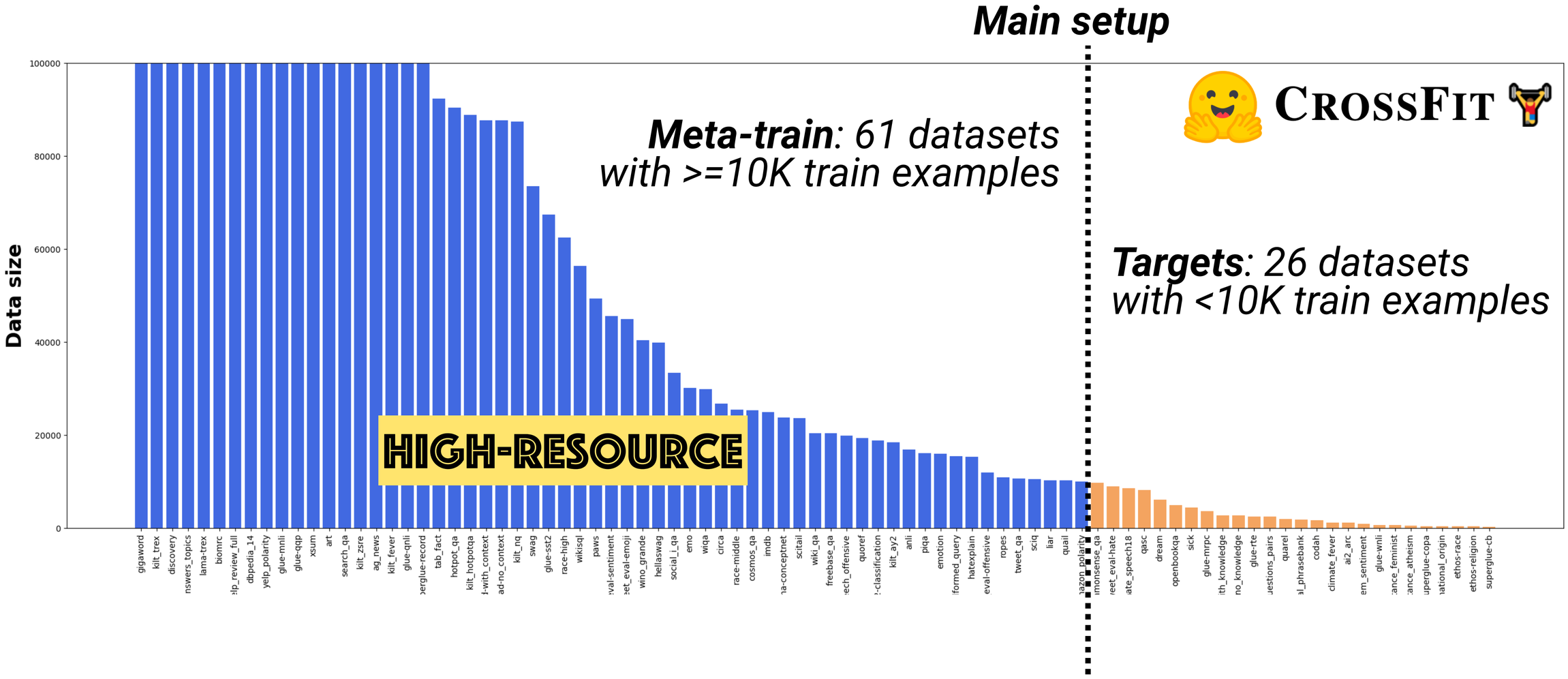


See the paper for more setups!

(Lhoest et al. 2021, Ye et al. 2021)

Experimental Setup

In total, experiment with **142** NLP datasets and **7** different metatrain-target splits

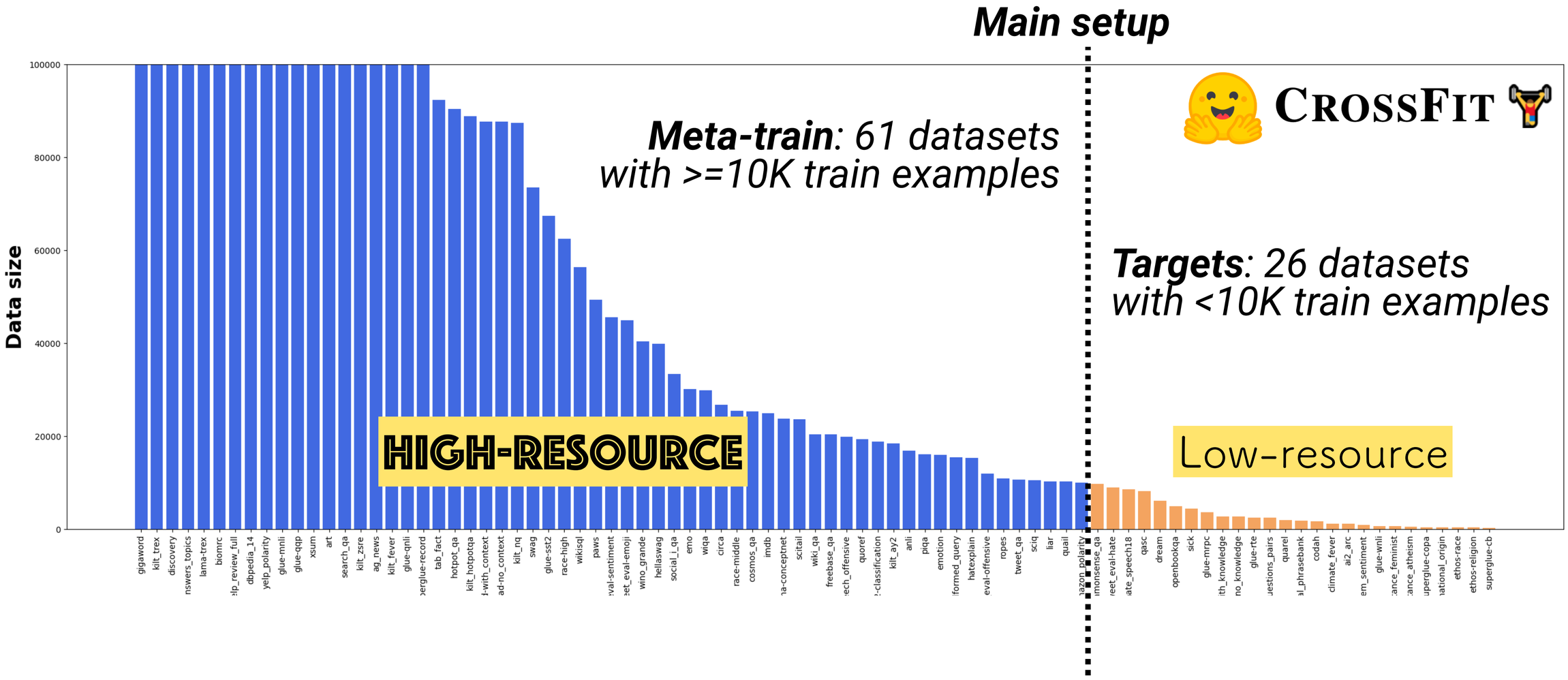


See the paper for more setups!

(Lhoest et al. 2021, Ye et al. 2021)

Experimental Setup

In total, experiment with **142** NLP datasets and **7** different metatrain-target splits

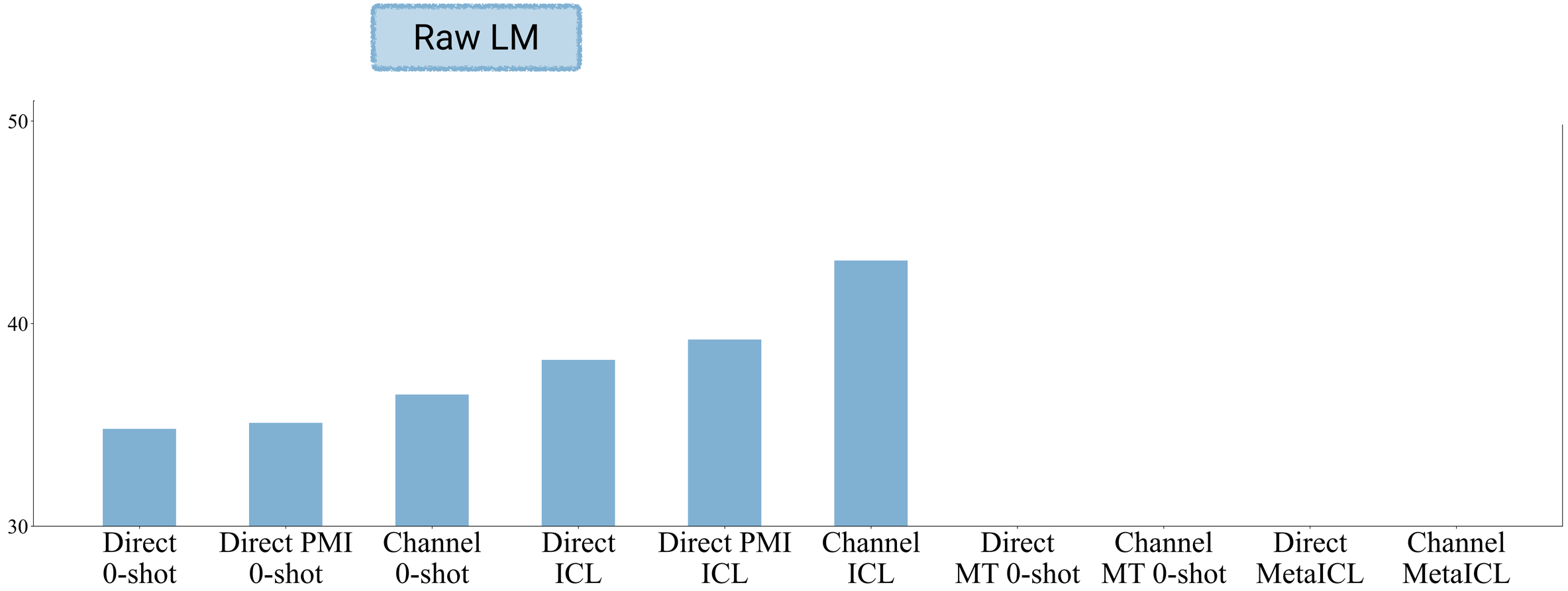


See the paper for more setups!

(Lhoest et al. 2021, Ye et al. 2021)

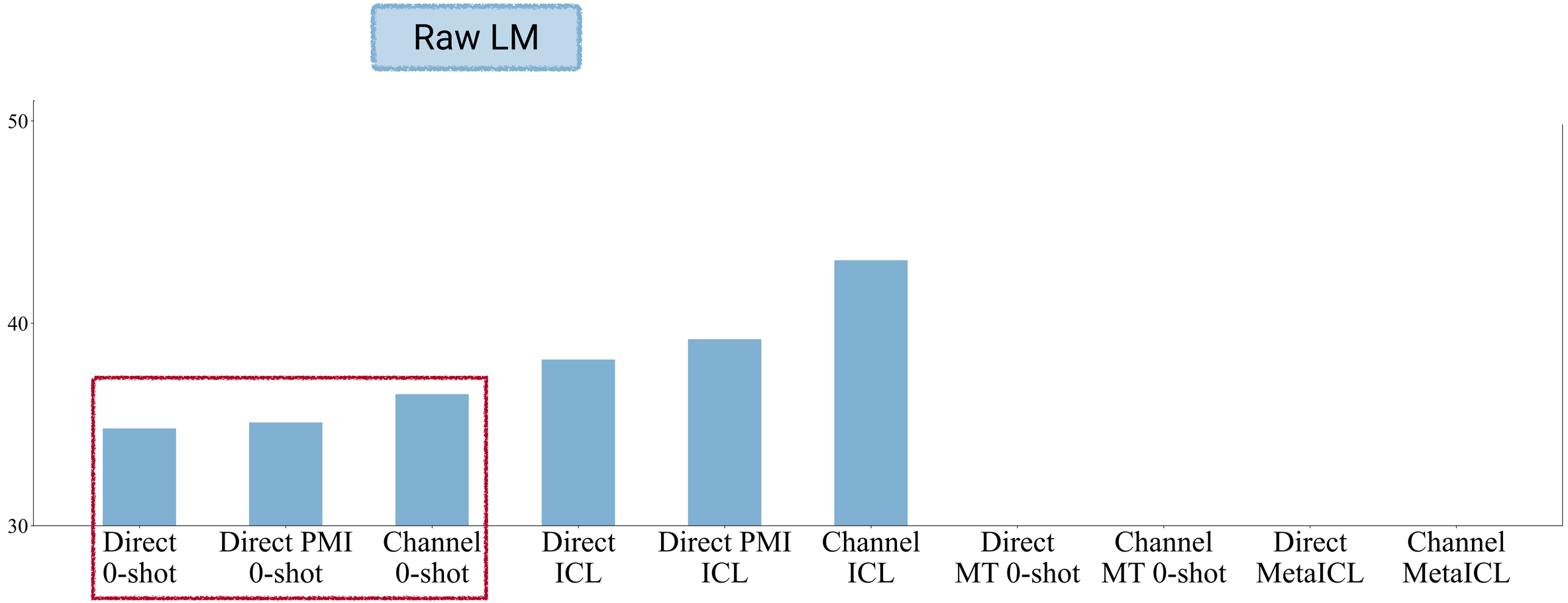
Base LM: GPT-2 Large

Results



Base LM: GPT-2 Large

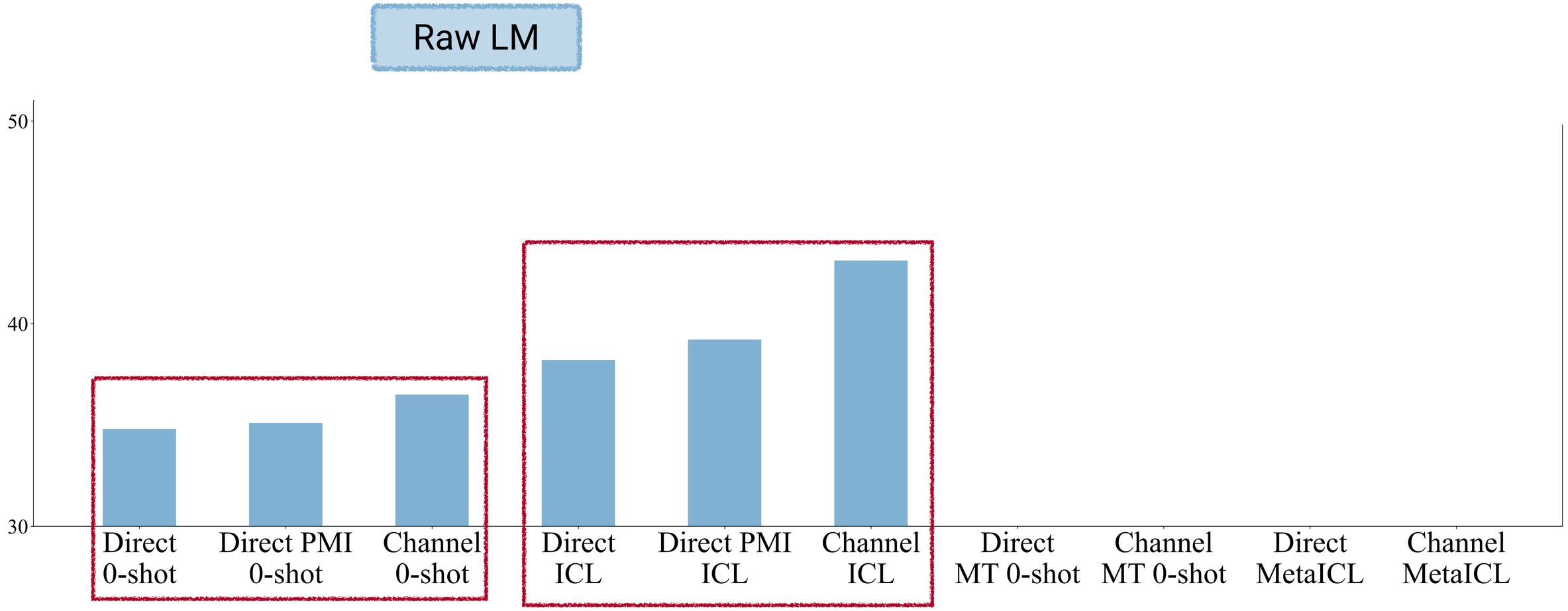
Results



** Direct: from Brown et al. 2020*
** Direct PMI: from Holtzman et al 2021, Zhao et al. 2021*
** Channel: from Min et al. 2022*

Base LM: GPT-2 Large

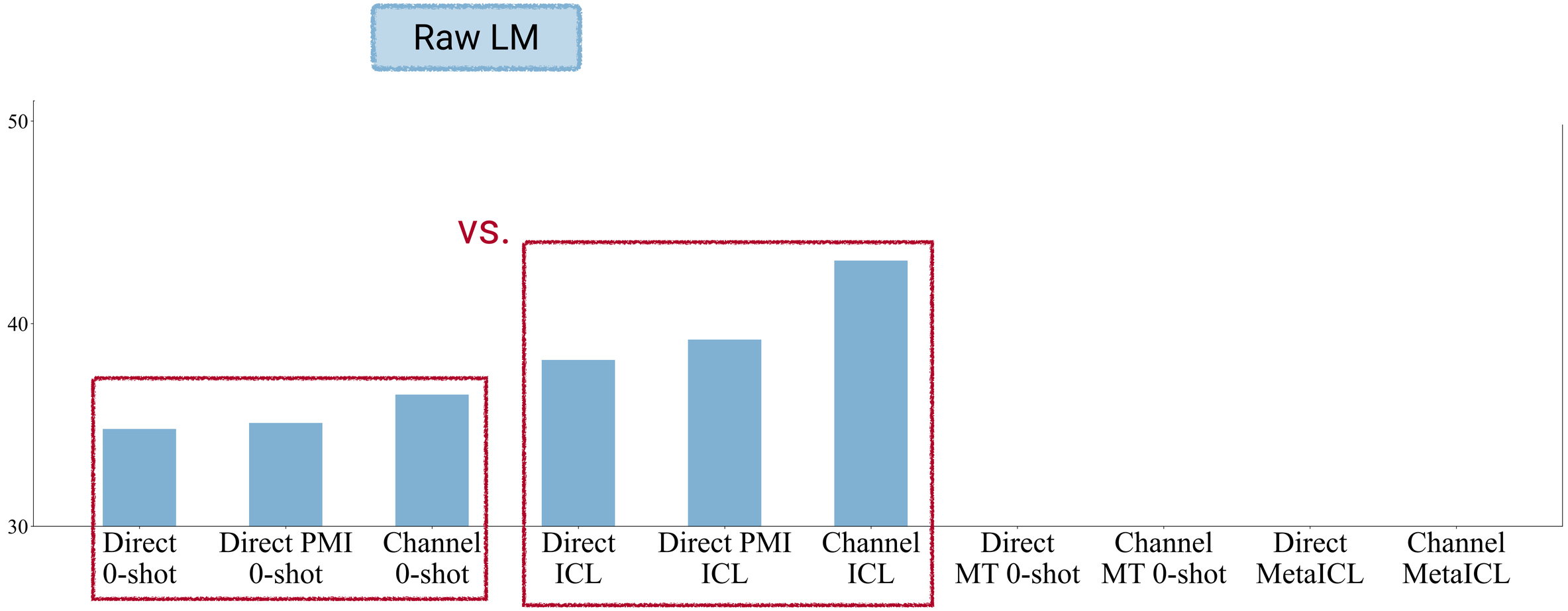
Results



** Direct: from Brown et al. 2020*
** Direct PMI: from Holtzman et al 2021, Zhao et al. 2021*
** Channel: from Min et al. 2022*

Base LM: GPT-2 Large

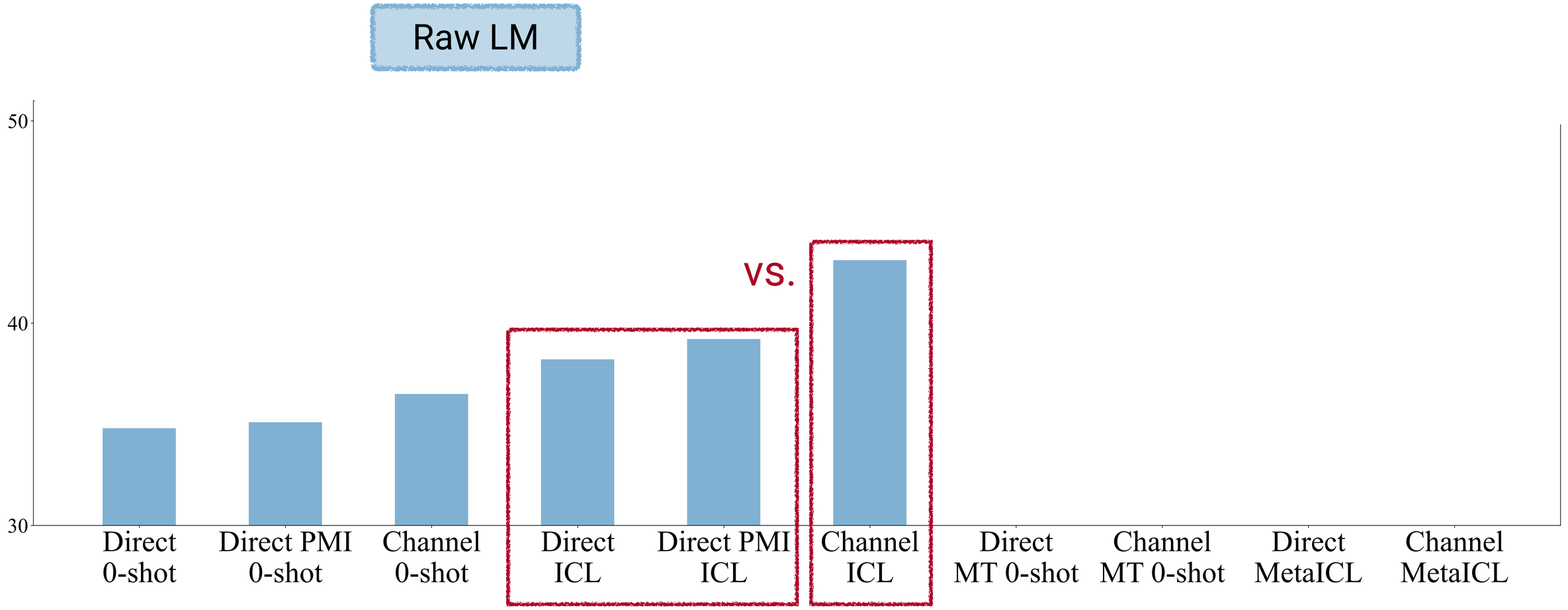
Results



** Direct: from Brown et al. 2020*
** Direct PMI: from Holtzman et al 2021, Zhao et al. 2021*
** Channel: from Min et al. 2022*

Base LM: GPT-2 Large

Results

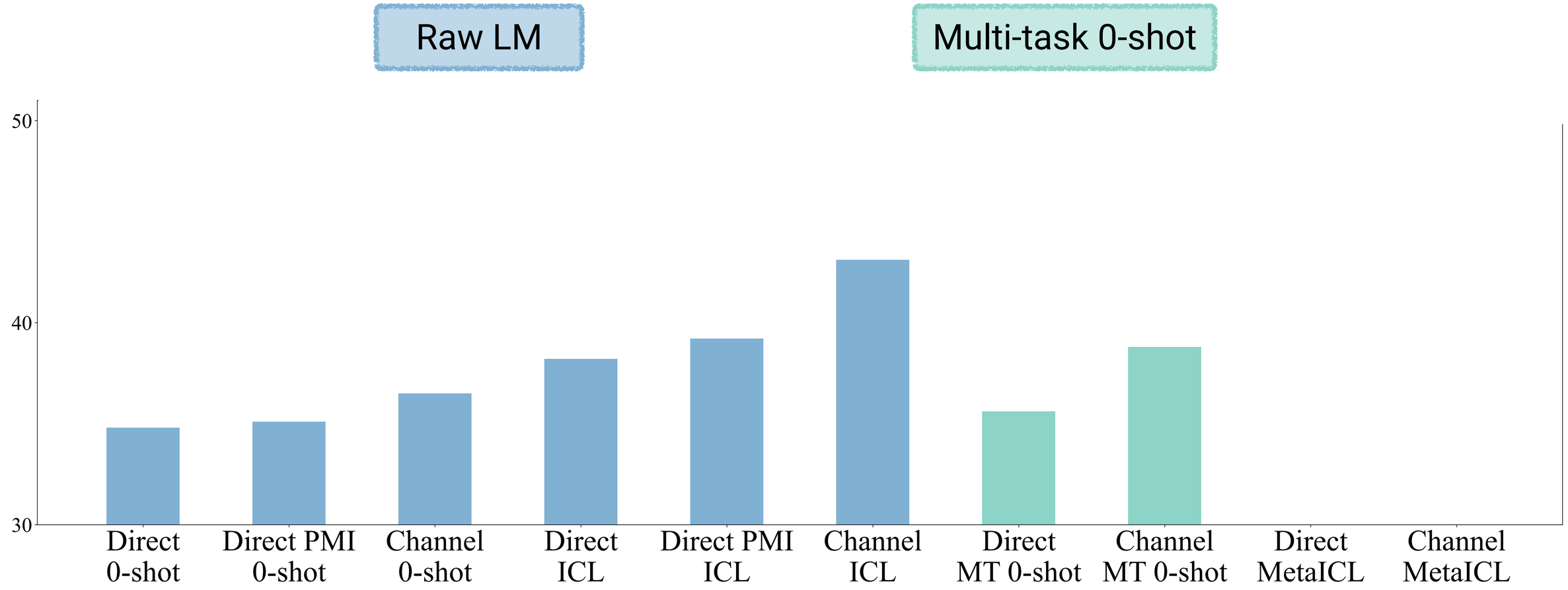


Channel ICL achieves the best result when no meta-training

Base LM: GPT-2 Large

Results

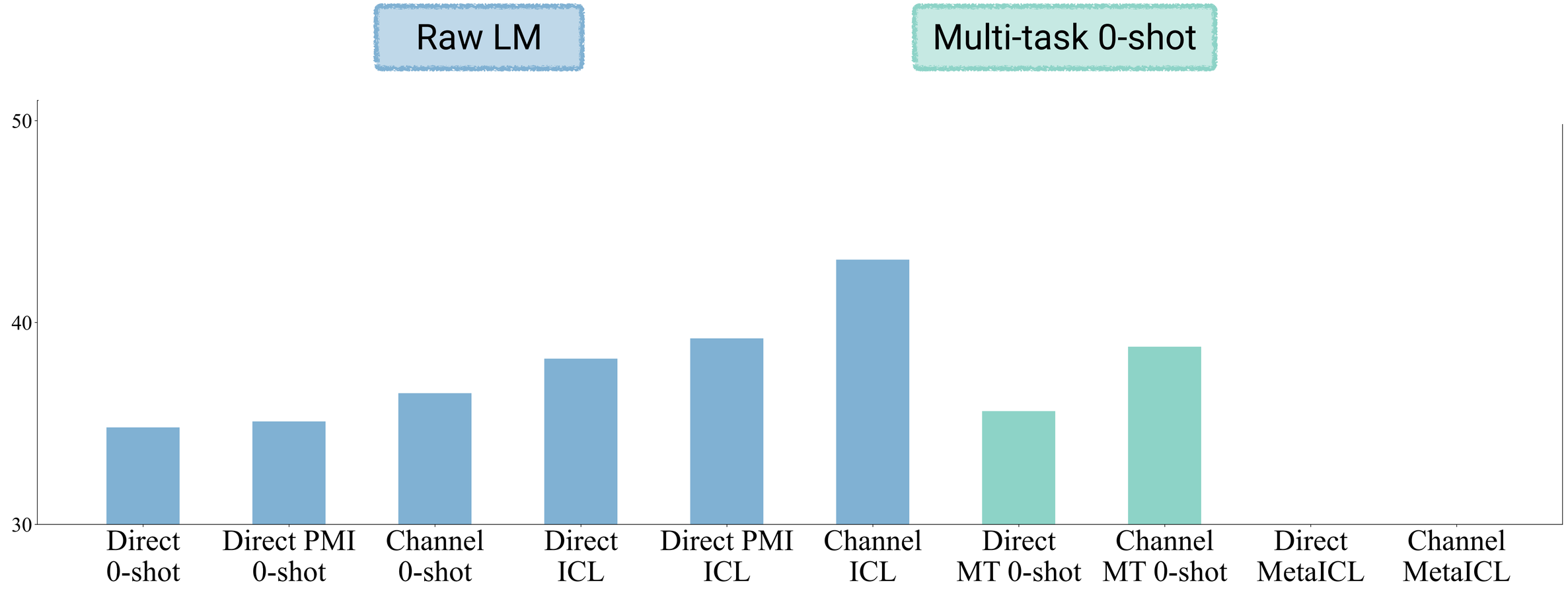
Multi-task training
but ***w/o*** in-context learning



Base LM: GPT-2 Large

Results

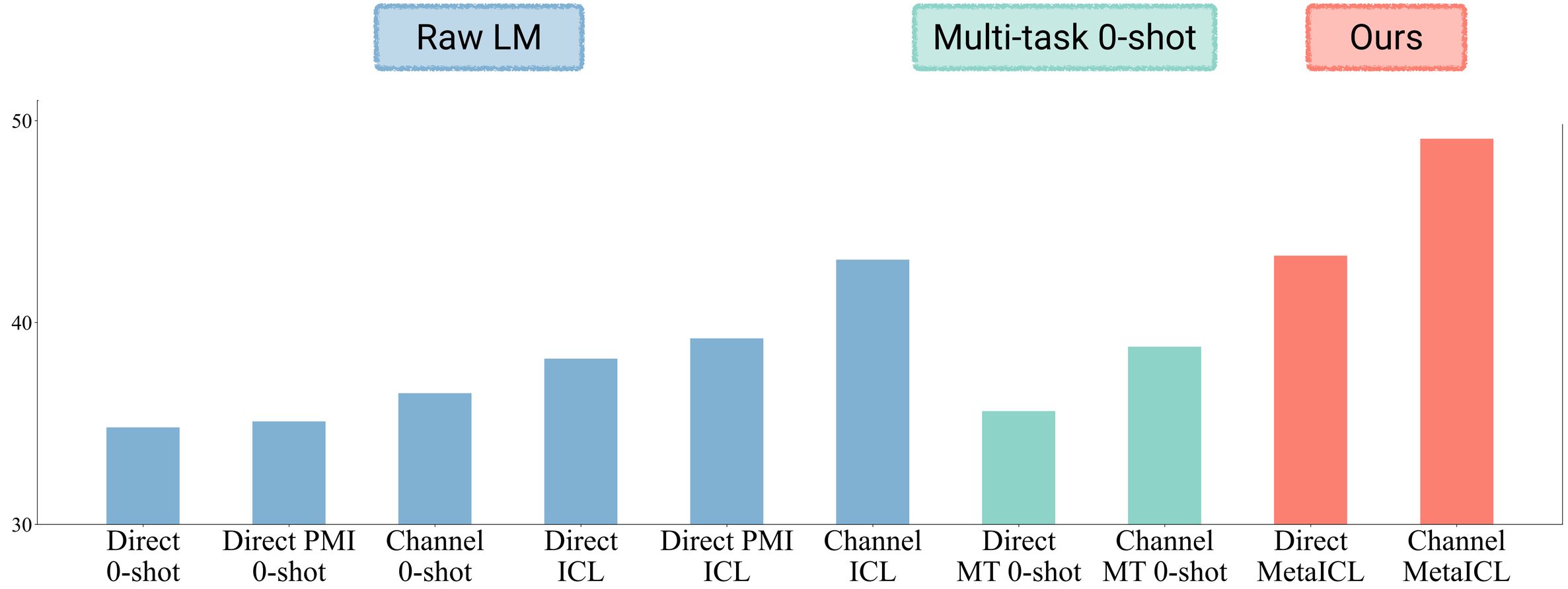
Multi-task training
but ***w/o*** in-context learning



Multi-task 0-shot is not as competitive

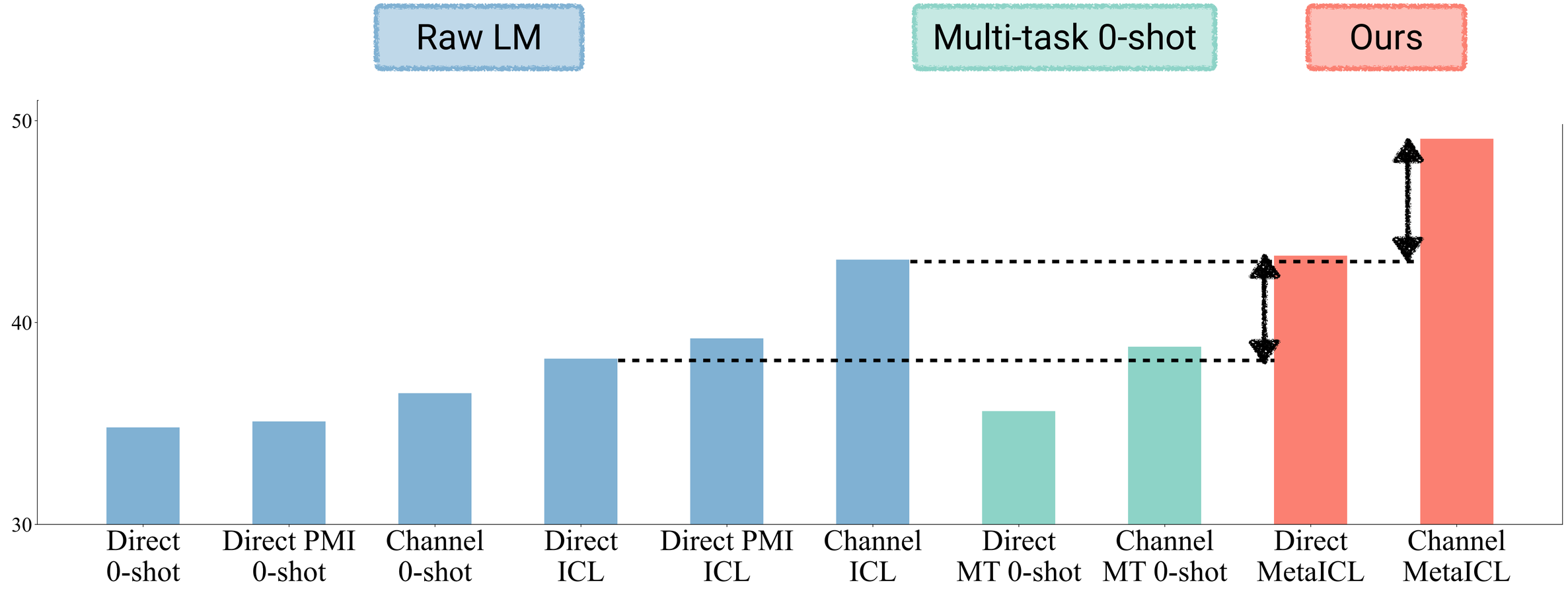
Base LM: GPT-2 Large

Results



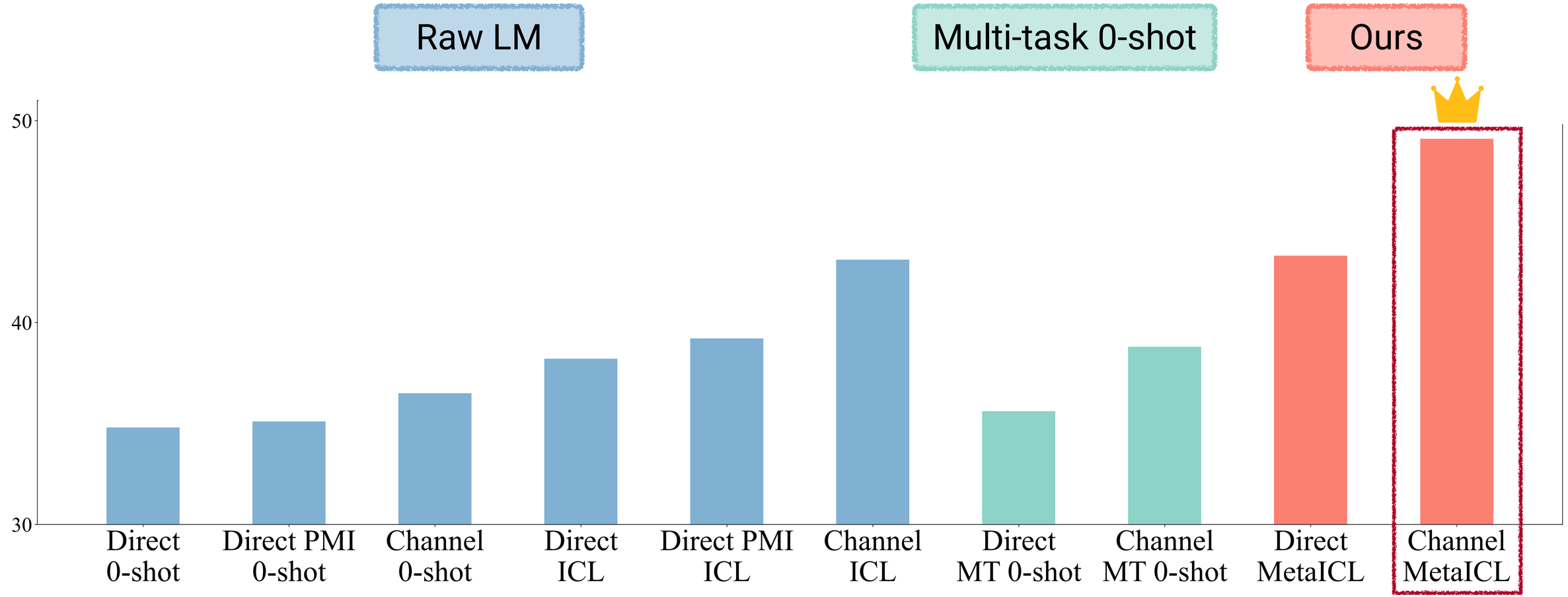
Base LM: GPT-2 Large

Results



Base LM: GPT-2 Large

Results

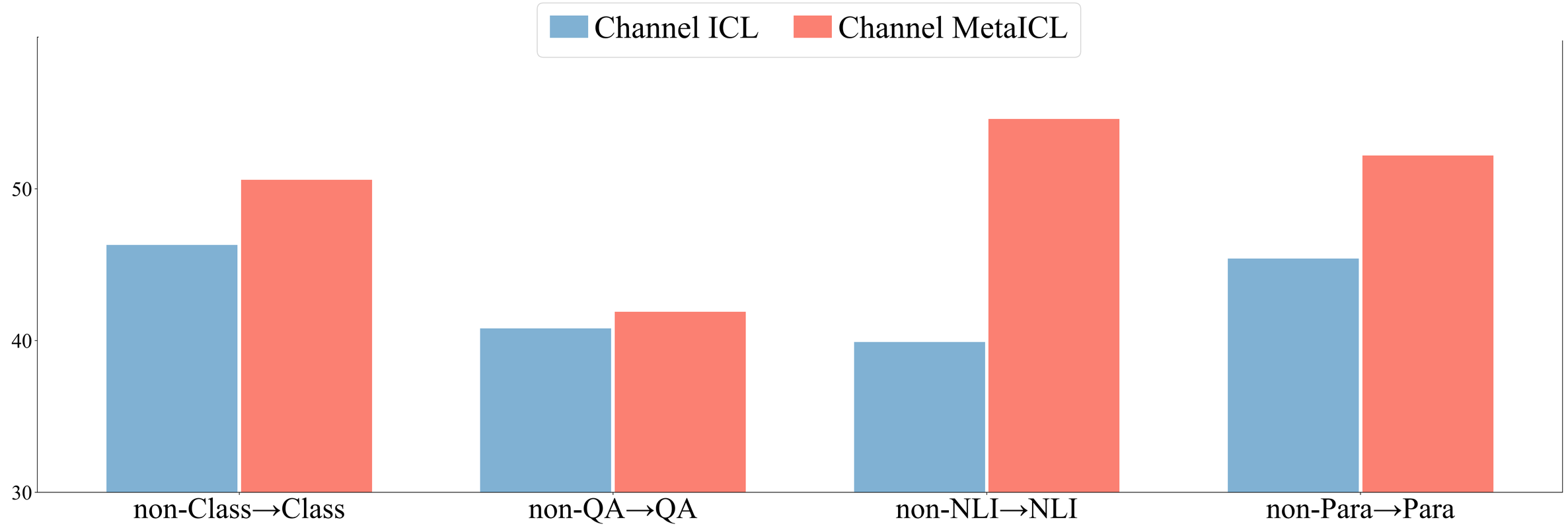


Channel MetaICL achieves the best result

Base LM: GPT-2 Large

Results

Other meta-training/targets splits

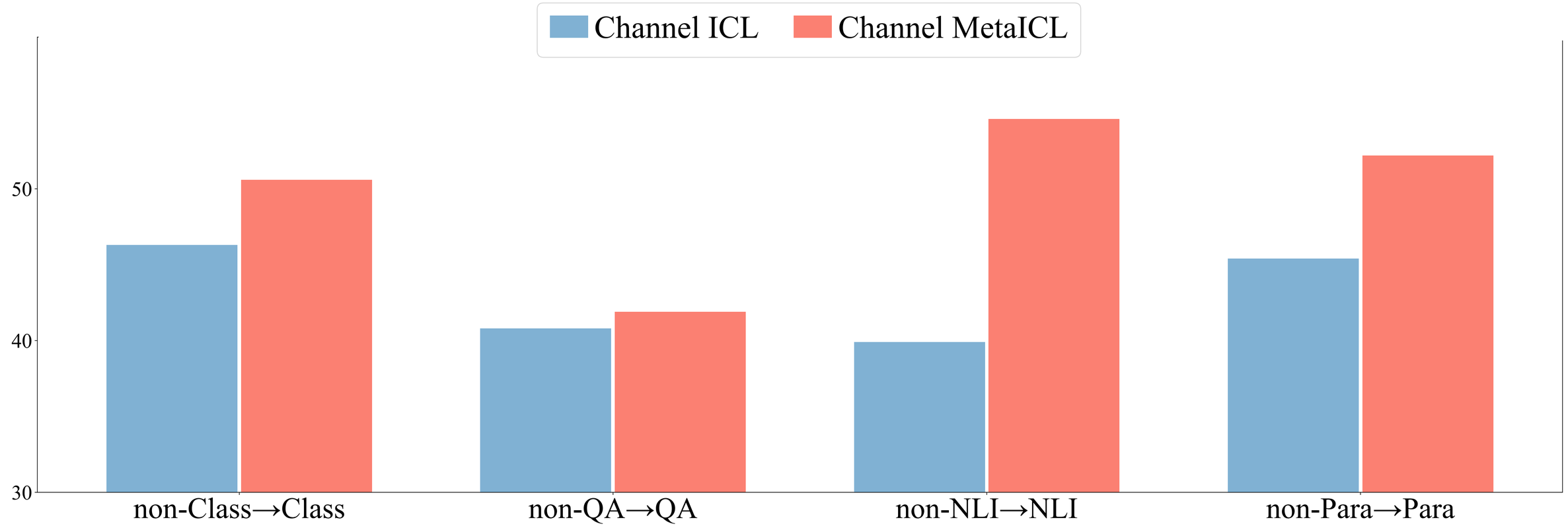


See the paper for more details!

Base LM: GPT-2 Large

Results

Other meta-training/targets splits



Large gains — even with no similar tasks at meta-training

See the paper for more details!

Base LM: GPT-2 Large

Results

vs. Instruction reading models

Base LM: GPT-2 Large

Results

vs. Instruction reading models

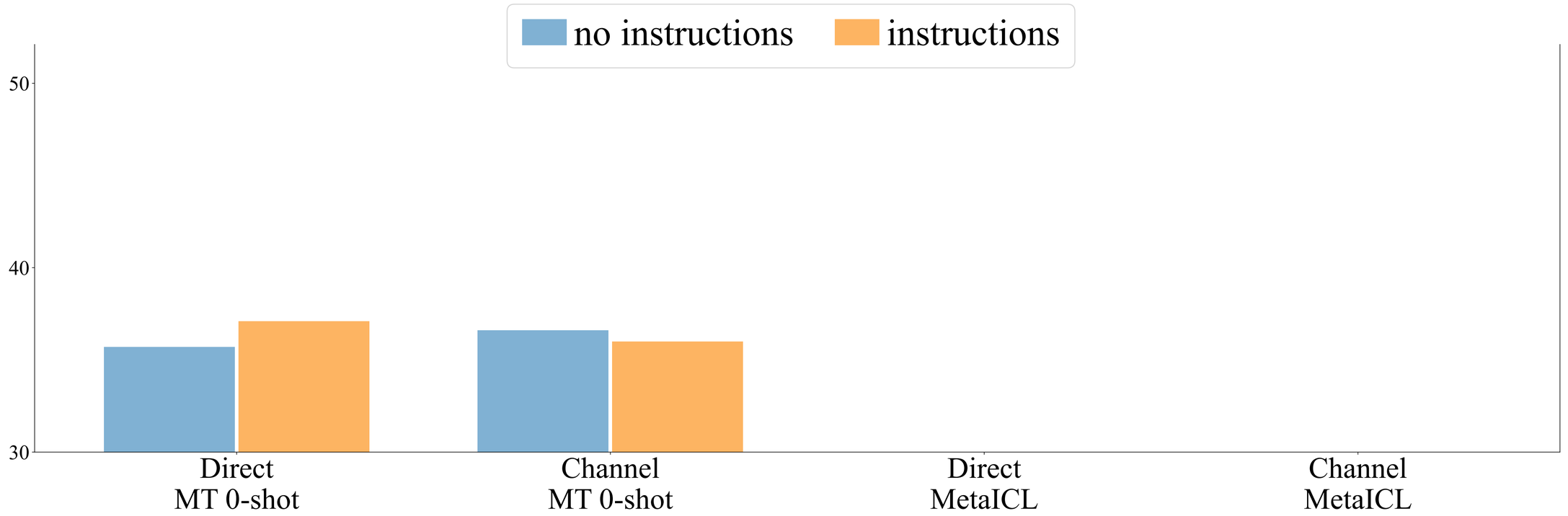
Used the data from T0 (San et al 2021) — 8.3 instructions per task

Base LM: GPT-2 Large

Results

vs. Instruction reading models

Used the data from T0 (San et al 2021) — 8.3 instructions per task



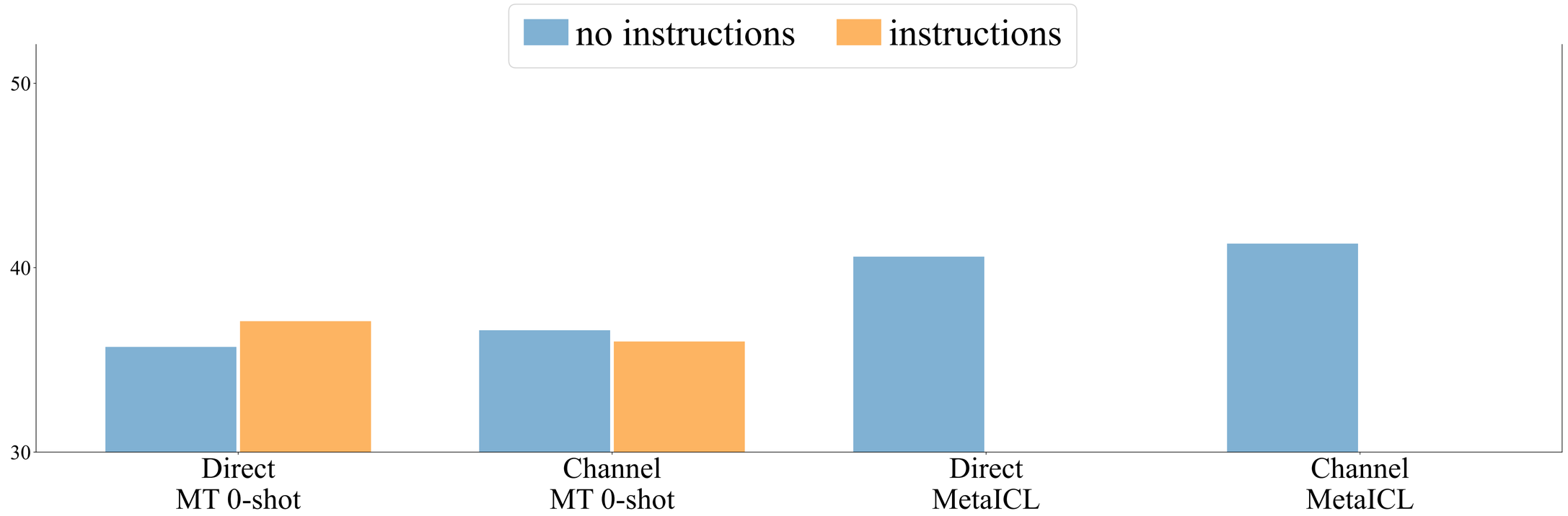
Multi-task model w/o in-context learning benefits from instructions

Base LM: GPT-2 Large

Results

vs. Instruction reading models

Used the data from T0 (San et al 2021) — 8.3 instructions per task



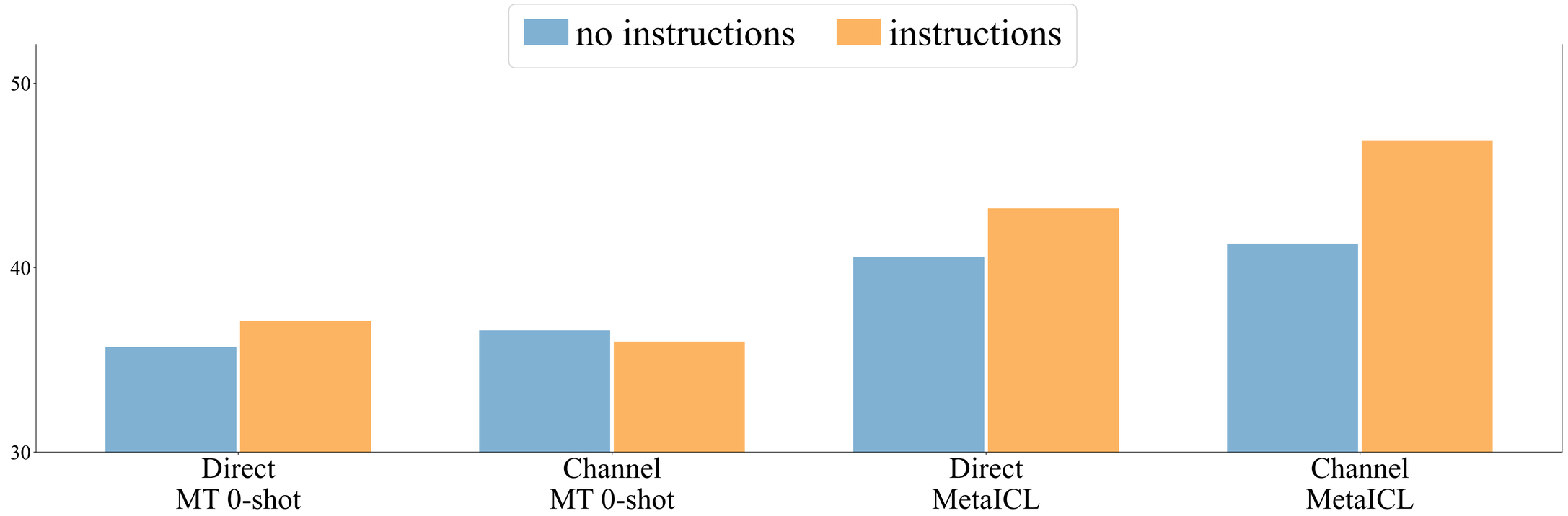
Training to in-context learn is better than training to read instructions

Base LM: GPT-2 Large

Results

vs. Instruction reading models

Used the data from T0 (San et al 2021) — 8.3 instructions per task

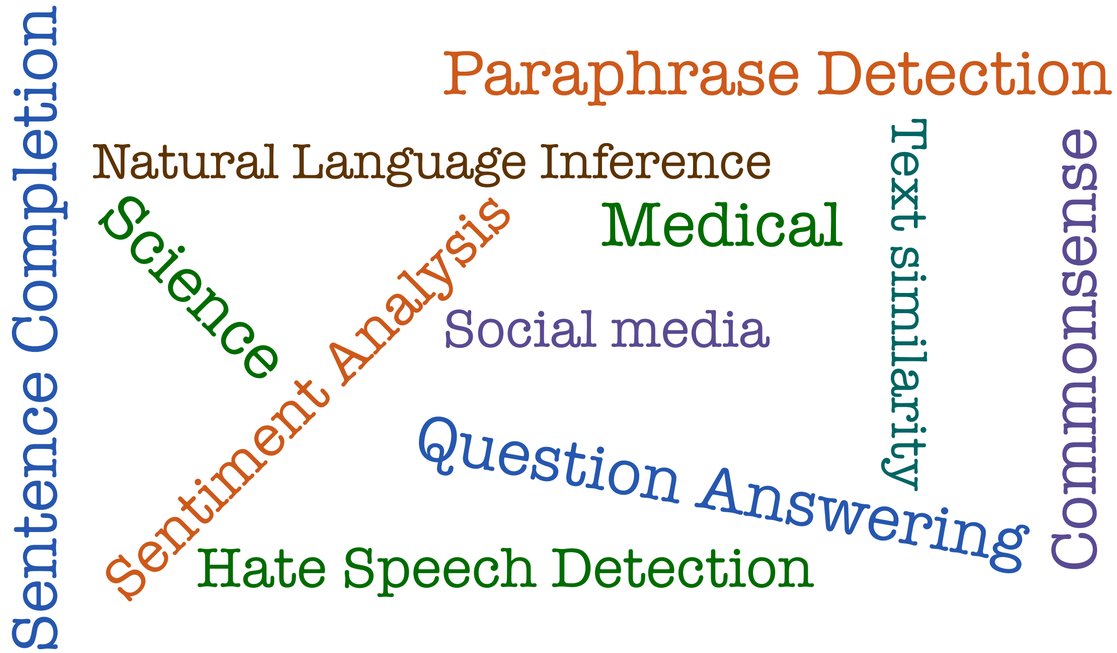


MetaICL + instructions reading achieves the best performance

Why does MetaICL work?

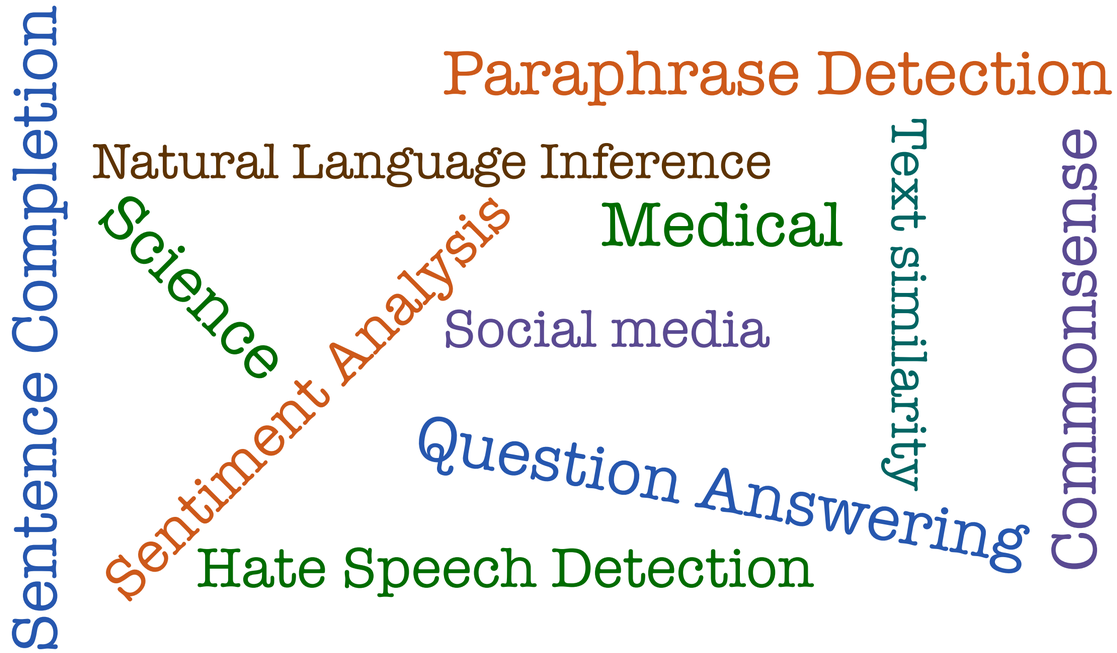
Why does MetaICL work?

1) The model benefits from *related* tasks during meta-training



Why does MetaICL work?

1) The model benefits from *related* tasks during meta-training

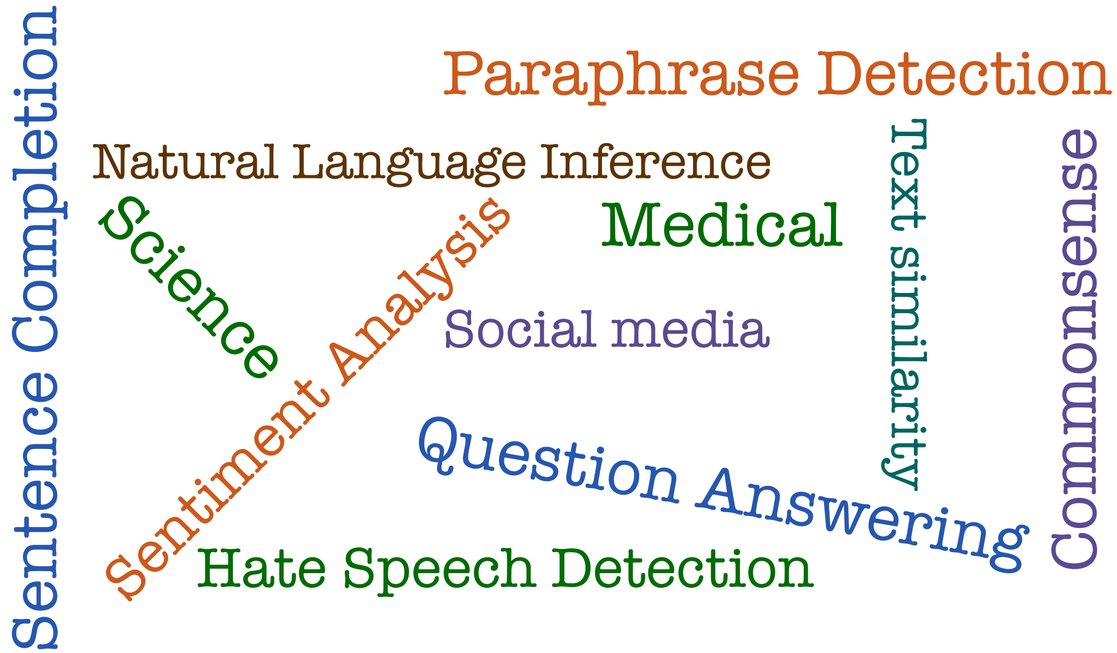


Already verified by a large body of work in multi-task learning

(Villa and Drissi 2002, Evgeniou and Pontil 2004, Ruder 2017, Aghajanyan et al. 2021)

Why does MetaICL work?

1) The model benefits from *related* tasks during meta-training



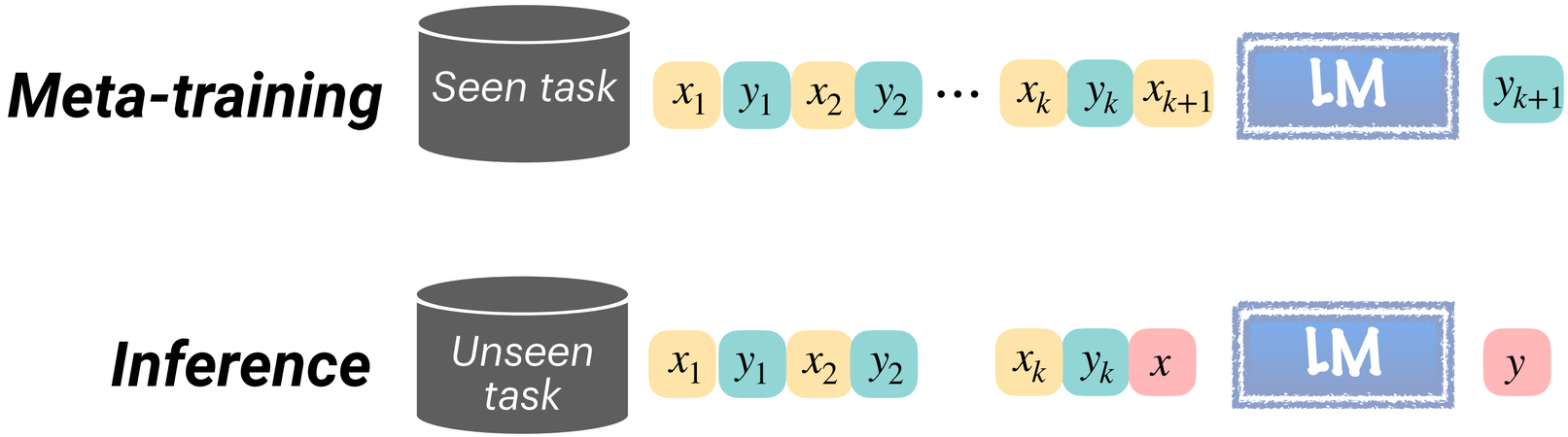
Already verified by a large body of work in multi-task learning

(Villa and Drissi 2002, Evgeniou and Pontil 2004, Ruder 2017, Aghajanyan et al. 2021)

However, Multi-task 0-shot is not enough

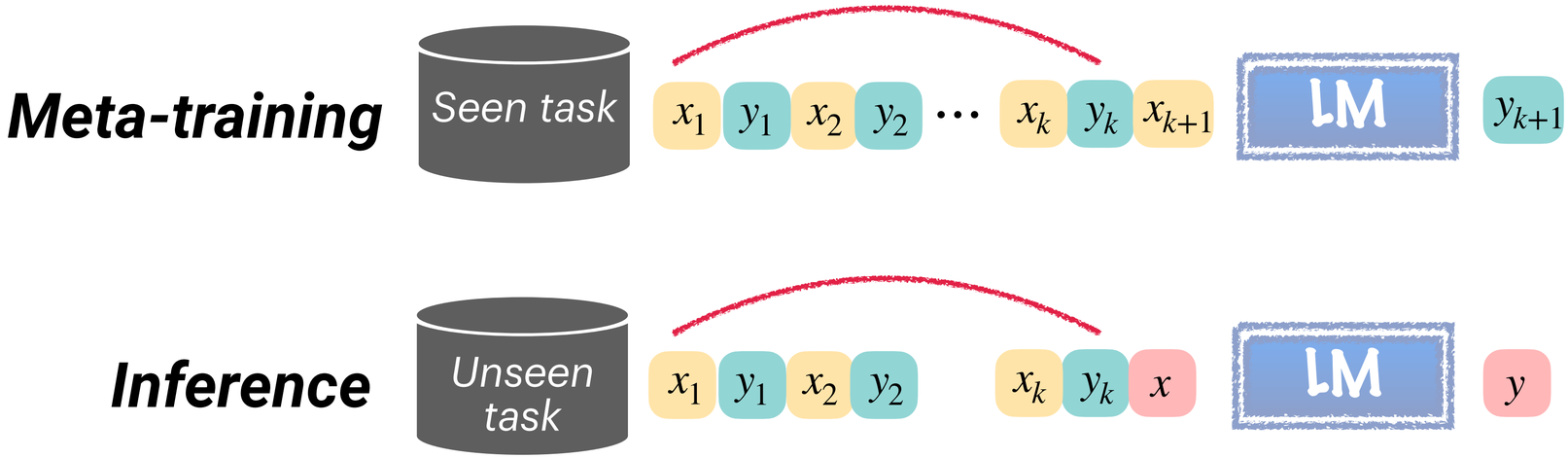
Why does MetaICL work?

2) The model learns the ability to *in-context learn a given task*



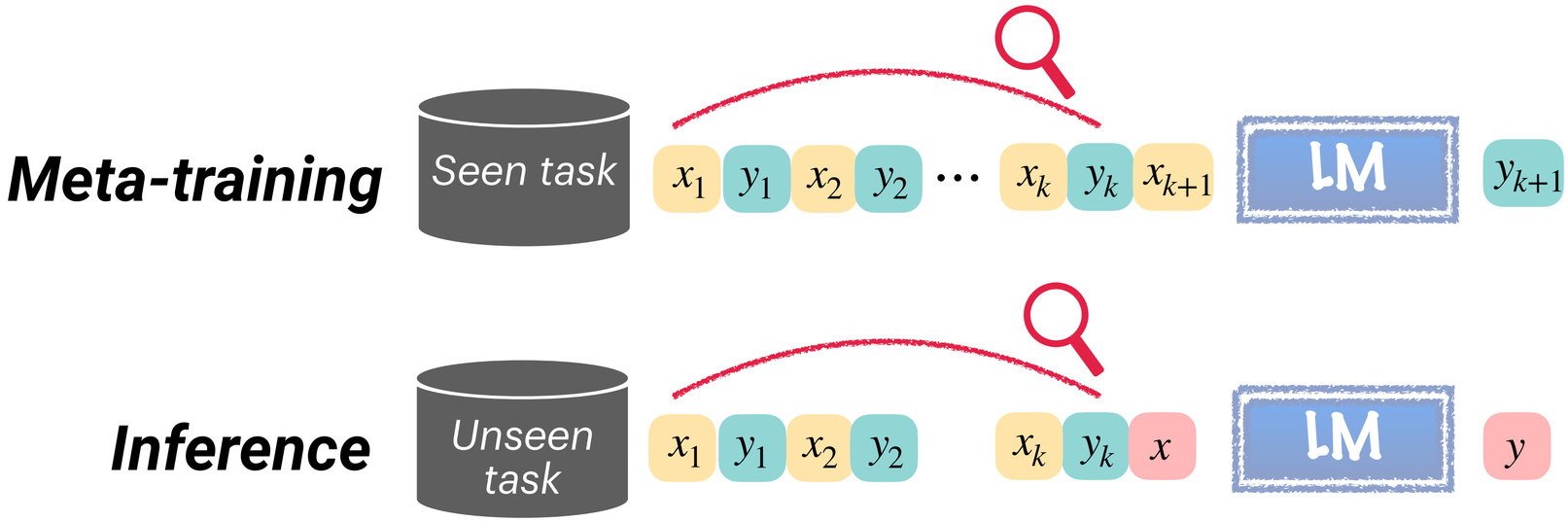
Why does MetaICL work?

2) The model learns the ability to *in-context learn a given task*



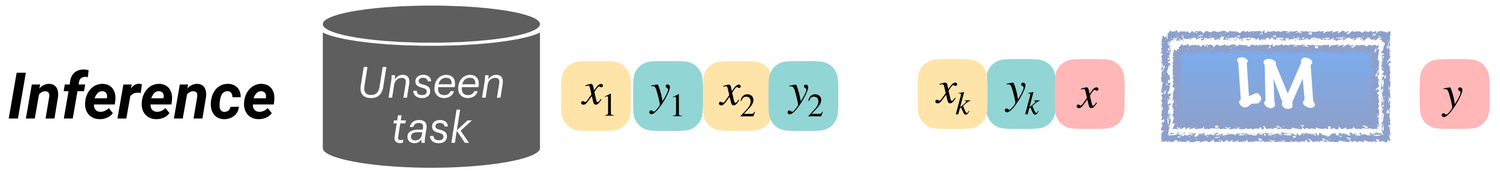
Why does MetaICL work?

2) The model learns the ability to *in-context learn a given task*



Why does MetaICL work?

2) The model learns the ability to *in-context learn a given task*

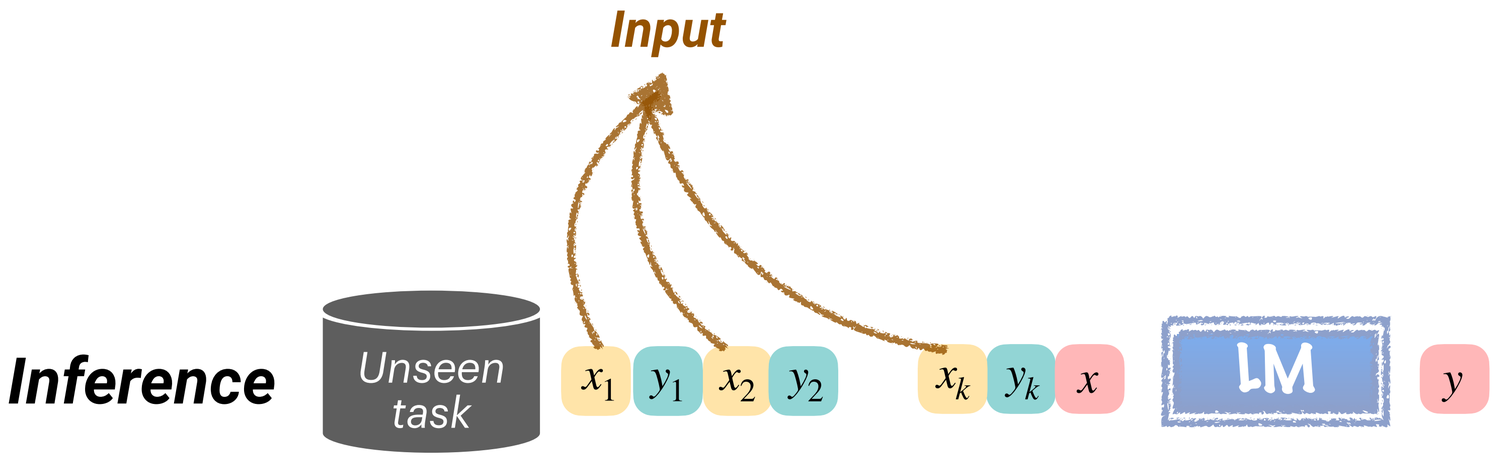


*"Meta-training encourages exploiting **input distribution** & **output space**,*

— Min et al. 2022. "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?"

Why does MetaICL work?

2) The model learns the ability to *in-context learn a given task*

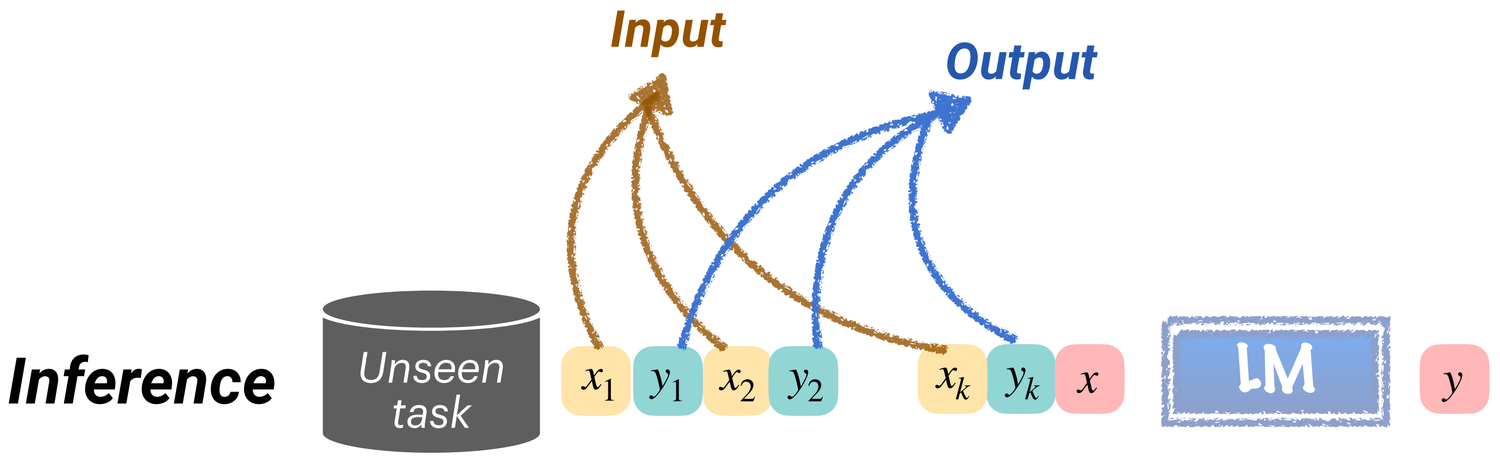


*"Meta-training encourages exploiting **input distribution** & **output space**,*

— Min et al. 2022. "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?"

Why does MetaICL work?

2) The model learns the ability to *in-context learn a given task*

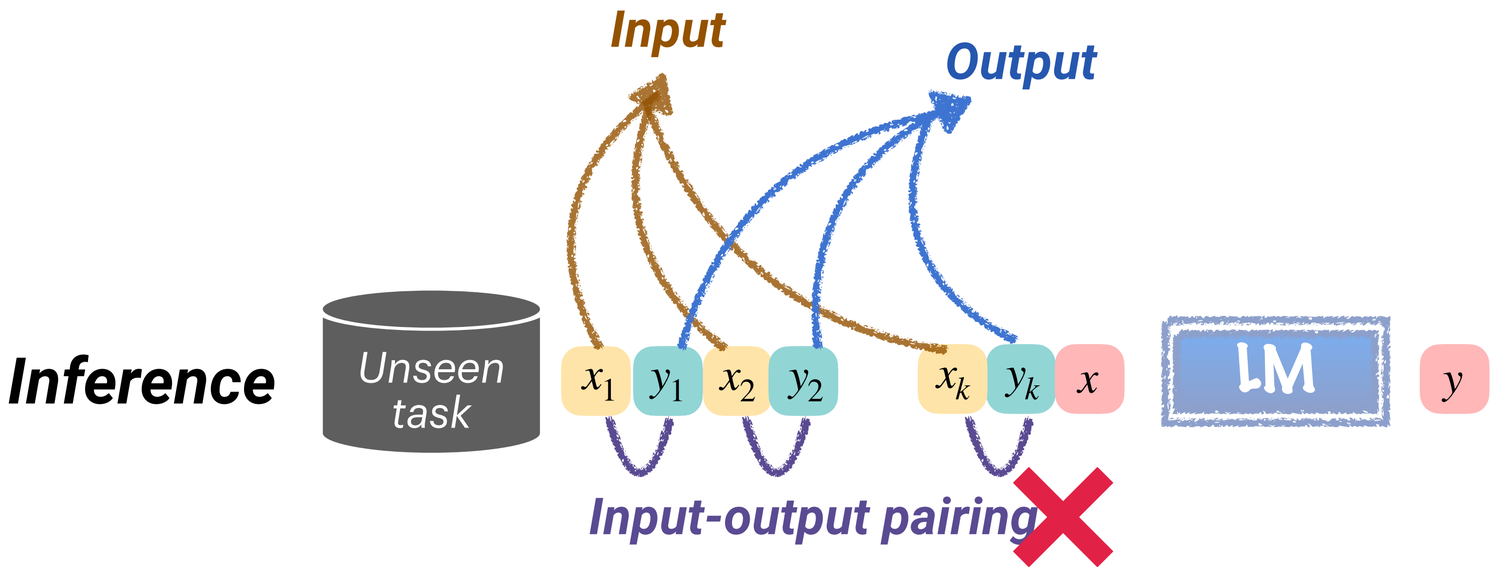


*"Meta-training encourages exploiting **input distribution** & **output space**,*

— Min et al. 2022. "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?"

Why does MetaICL work?

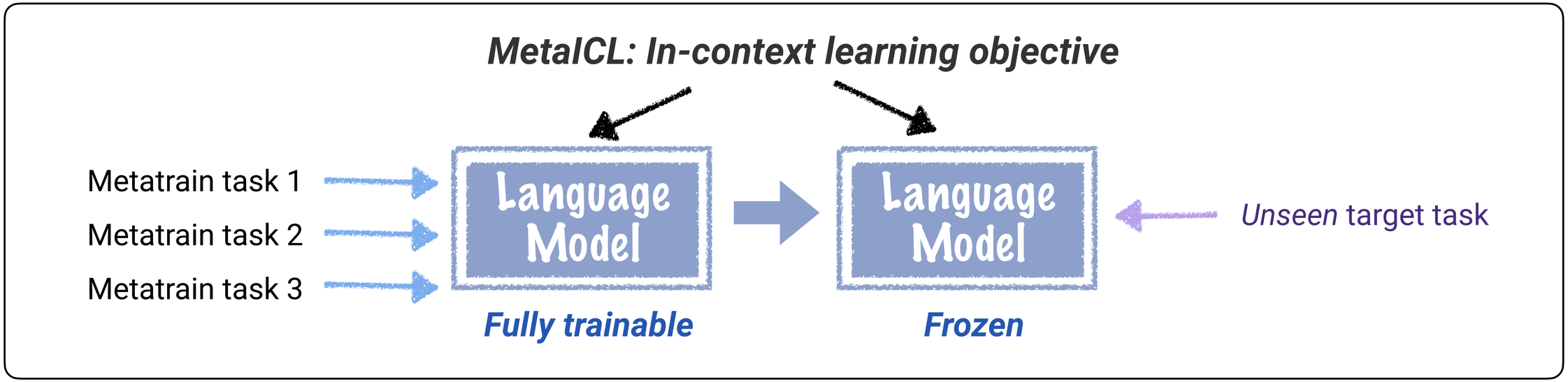
2) The model learns the ability to *in-context learn a given task*



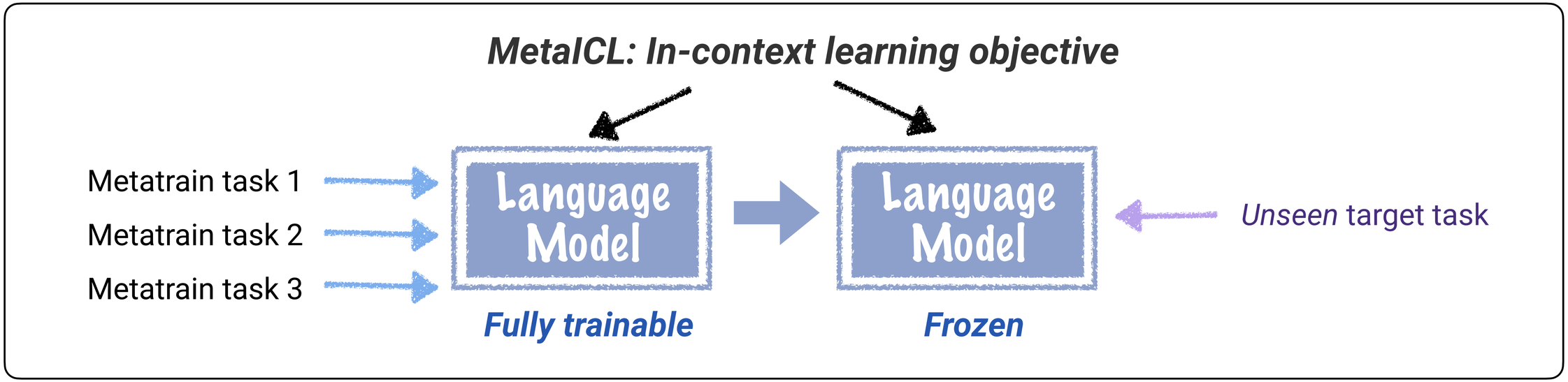
*"Meta-training encourages exploiting **input distribution** & **output space**, and ignoring **input-output pairing**"*

— Min et al. 2022. "Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?"

Takeaways


MetaICL: In-context learning objective
Metatrain task 1
Metatrain task 2
Metatrain task 3
Language Model
Language Model
Unseen target task
Fully trainable
Frozen

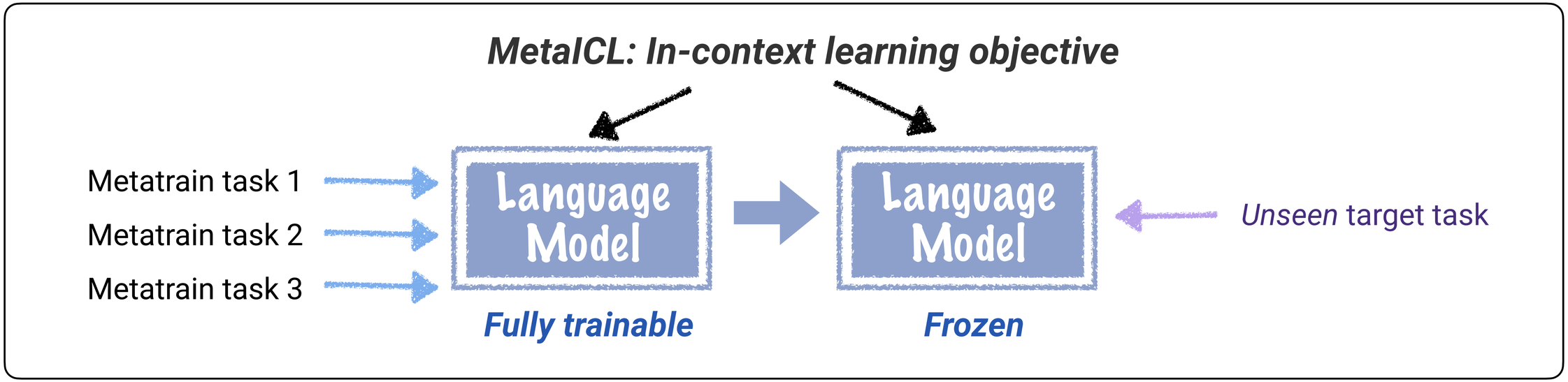
Takeaways



On seven different splits

★ Significantly outperform strong no-metatrained baseline

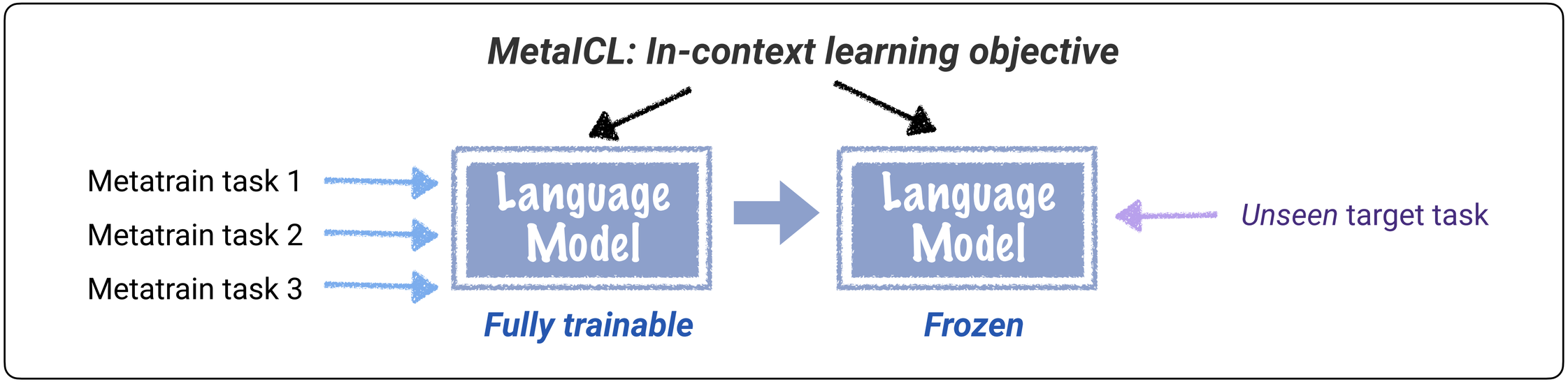
Takeaways



On seven different splits

- ★ Significantly outperform strong no-metatrained baseline
- ★ Consistent gains even w/no overlap in task categories

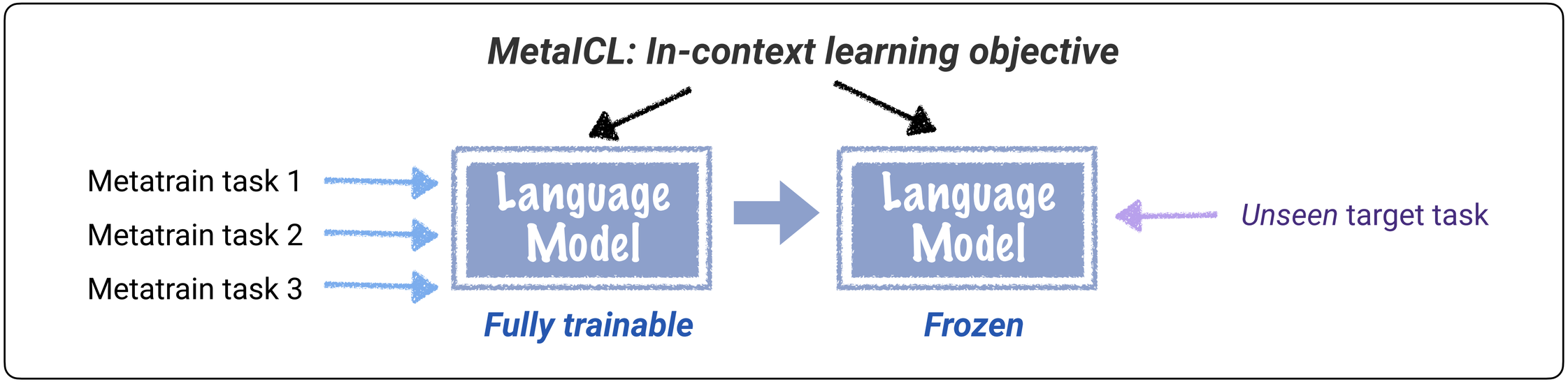
Takeaways



On seven different splits

- ★ Significantly outperform strong no-metatrained baseline
- ★ Consistent gains even w/no overlap in task categories
- ★ Outperforms 8x bigger model

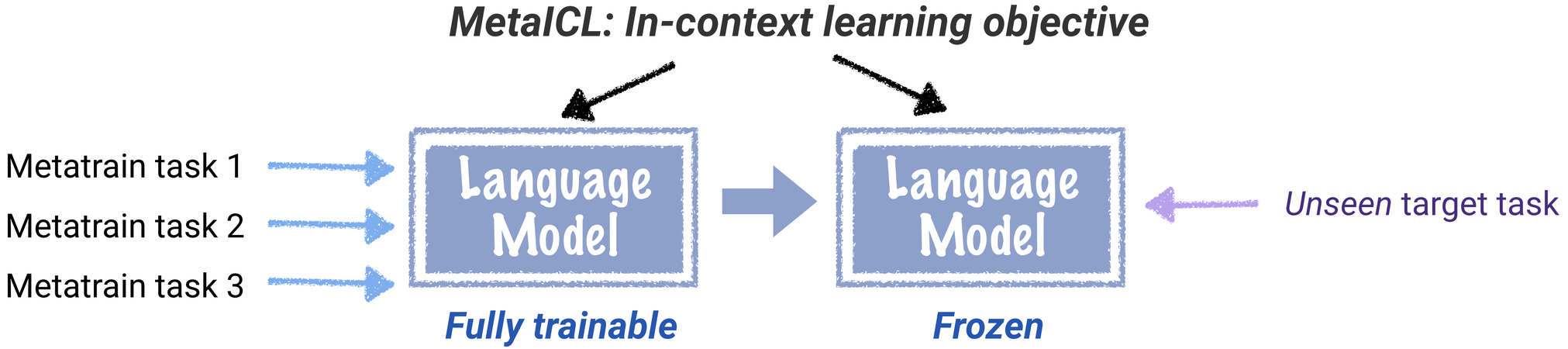
Takeaways



On seven different splits

- ★ Significantly outperform strong no-metatrained baseline
- ★ Consistent gains even w/no overlap in task categories
- ★ Outperforms 8x bigger model
- ★ Complementary to natural instructions

Takeaways

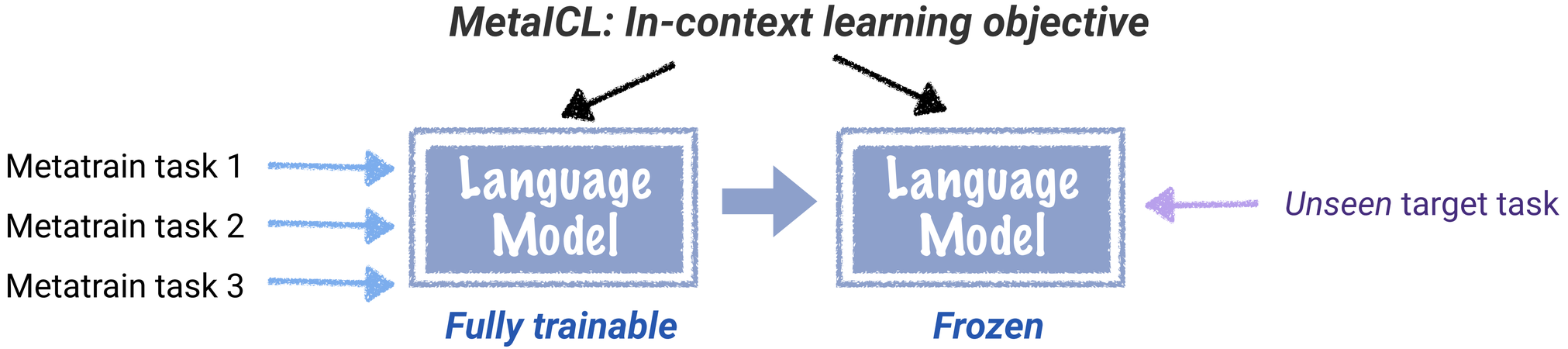


On seven different splits

- ★ Significantly outperform strong no-metatrained baseline
- ★ Consistent gains even w/no overlap in task categories
- ★ Outperforms 8x bigger model
- ★ Complementary to natural instructions

- ◎ Limited to classification/multi-choice

Takeaways

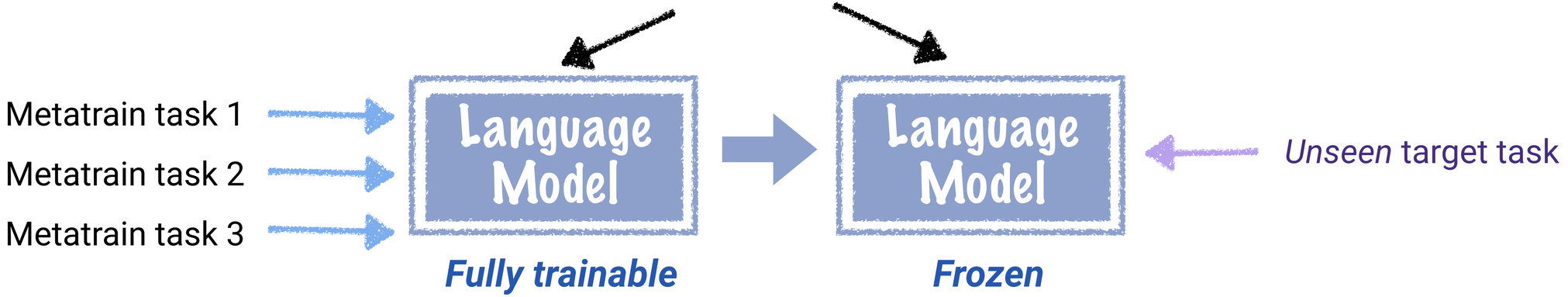


On seven different splits

- ★ Significantly outperform strong no-metatrained baseline
- ★ Consistent gains even w/no overlap in task categories
- ★ Outperforms 8x bigger model
- ★ Complementary to natural instructions

- ◎ Limited to classification/multi-choice
- ◎ Still lags behind metatrain→finetune

Takeaways



On seven different splits

- ★ Significantly outperform strong no-metatrained baseline
- ★ Consistent gains even w/no overlap in task categories
- ★ Outperforms 8x bigger model
- ★ Complementary to natural instructions

- ◎ Limited to classification/multi-choice
- ◎ Still lags behind metatrain→finetune
- ◎ How to ensure the model refers to input-output pairing?

Demo: http://qa.cs.washington.edu:2021

Few-shot Learning through Channel MetaICL

Sewon Min, Mike Lewis, Luke Zettlemoyer, Hannaneh Hajishirzi. 2021. "MetaICL: Learning to Learn In Context"
Sewon Min, Mike Lewis, Hannaneh Hajishirzi, Luke Zettlemoyer. 2021. "Noisy Channel Language Model Prompting for Few-Shot Text Classification"
Contact/bug report: Sewon Min (,)
[Show Me Details!]

Instruction:

Give the model 16 training examples (by writing yourself or choosing from an example task) and the test input, and see what model returns!
You do not need to use any templates or instructions to tell what the task is.

Channel LM Channel MetaICL

Recommended Tasks

=====Your Own Task=====

Get Random Examples!

Training examples

Input: Our model outperforms a range of competitive baselines.

Output: positive

Input: Our model achieves the new state-of-the-art.

Test Input

Write your own input

Run

Please enter the test input

Demo: http://qa.cs.washington.edu:2021

Few-shot Learning through Channel MetaICL

Sewon Min, Mike Lewis, Luke Zettlemoyer, Hannaneh Hajishirzi. 2021. "MetaICL: Learning to Learn In Context"
Sewon Min, Mike Lewis, Hannaneh Hajishirzi, Luke Zettlemoyer. 2021. "Noisy Channel Language Model Prompting for Few-Shot Text Classification"
Contact/bug report: Sewon Min (,)
[Show Me Details!]

Instruction:

Give the model 16 training examples (by writing yourself or choosing from an example task) and the test input, and see what model returns!
You do not need to use any templates or instructions to tell what the task is.

Channel LM Channel MetaICL

Recommended Tasks

=====Your Own Task=====

Get Random Examples!

Training examples

Input: Our model outperforms a range of competitive baselines.

Output: positive

Input: Our model achieves the new state-of-the-art.

Test Input

Write your own input

Run

Please enter the test input

Thank you for listening

Paper: arxiv.org/abs/2110.15943
Code: github.com/facebookresearch/MetaICL
Demo: http://qa.cs.washington.edu:2021/

Contact: sewon@cs.washington.edu / @sewon__min

UW NLP Meta AI2 NAACL 2022