Meta-Learning to Compositionally Generalize

Henry Conklin[1], Bailin Wang[1*], Kenny Smith[1], Ivan Titov[1,2]*

[1]University of Edinburgh [2]University of Amsterdam [*]equal contribution

the meaning of a sentence is constructed from the meaning of its parts and the way in which they are combined

(Cann, 1993)

Compositionality

Generalization

Enables robust generalization outside of prior experience

Compositionality

Generalization

Enables robust generalization outside of prior experience

```
The deer ran across the road last night.
```

Compositionality

Generalization

Enables robust generalization outside of prior experience

```
The deer ran across the road last night.

The chicken walked into town.
```

Compositionality

Generalization

Enables robust generalization outside of prior experience

```
The deer ran across the road last night.

The chicken walked into town.
```

```
The deer walked into town.
```

Compositionality

Generalization

Enables robust generalization outside of prior experience

```
The deer ran across the road last night.

The chicken walked into town.


I don't like pears, I find them sinister.
```

```
The deer walked into town.
```

Compositionality

Generalization

Enables robust generalization outside of prior experience

```
The deer ran across the road last night.

The chicken walked into town.


I don’t like pears, I find them sinister.

apples
```

```
The deer walked into town.
```

Compositionality

Generalization

Enables robust generalization outside of prior experience

The deer ran across the road last night.

The chicken walked into town.

I don’t like pears, I find them sinister.

apples

The deer walked into town.

I don’t like apples, I find them sinister.

Compositionality

Current Limitations

State of the art neural models struggle to generalize outside of their training distribution

Compositionality

Current Limitations

State of the art neural models struggle to generalize outside of their training distribution

- They struggle to use new words in a compositional context (Lake and Baroni, 2018)

Compositionality

Current Limitations

State of the art neural models struggle to generalize outside of their training distribution

- They struggle to use new words in a compositional context (Lake and Baroni, 2018)
- Difficulty interpreting known words in new contexts (Keysers et al., 2020)

Compositionality

Current Limitations

State of the art neural models struggle to generalize outside of their training distribution

- They struggle to use new words in a compositional context (Lake and Baroni, 2018)
- Difficulty interpreting known words in new contexts (Keysers et al., 2020)
- Issues generalizing known words to new syntactic structures (Kim and Linzen, 2020)

Compositionality

Current Limitations

State of the art neural models struggle to generalize outside of their training distribution

Compositionality

Current Limitations

State of the art neural models struggle to generalize outside of their training distribution

- Models may prefer memorization over generalization (Liška et al., 2018)

Compositionality

Current Limitations

State of the art neural models struggle to generalize outside of their training distribution

- Models may prefer memorization over generalization (Liška et al., 2018)
- Models may memorize sections of their input (Hupkes et al., 2019)

Compositionality

Current Limitations

State of the art neural models struggle to generalize outside of their training distribution

- Models may prefer memorization over generalization (Liška et al., 2018)
- Models may memorize sections of their input (Hupkes et al., 2019)
- Limited memory may be key to why humans arrive at robust solutions (Griffiths, 2020)

Compositionality

Current Limitations

State of the art neural models struggle to generalize outside of their training distribution

- Models may prefer memorization over generalization (Liška et al., 2018)
- Models may memorize sections of their input (Hupkes et al., 2019)
- Limited memory may be key to why humans arrive at robust solutions (Griffiths, 2020)

Can we inhibit these models' ability to memorize?

Explicitly testing compositionality

COGS (Kim and Linzen, 2020)

Explicitly testing compositionality

COGS (Kim and Linzen, 2020)

```
The sailor dusted a boy .  →  * sailor ( x _ 1 ) ; dust . agent ( x _ 2 , x _ 1 ) AND
                              dust . theme ( x _ 2 , x _ 4 ) AND boy ( x _ 4 )
```

Explicitly testing compositionality

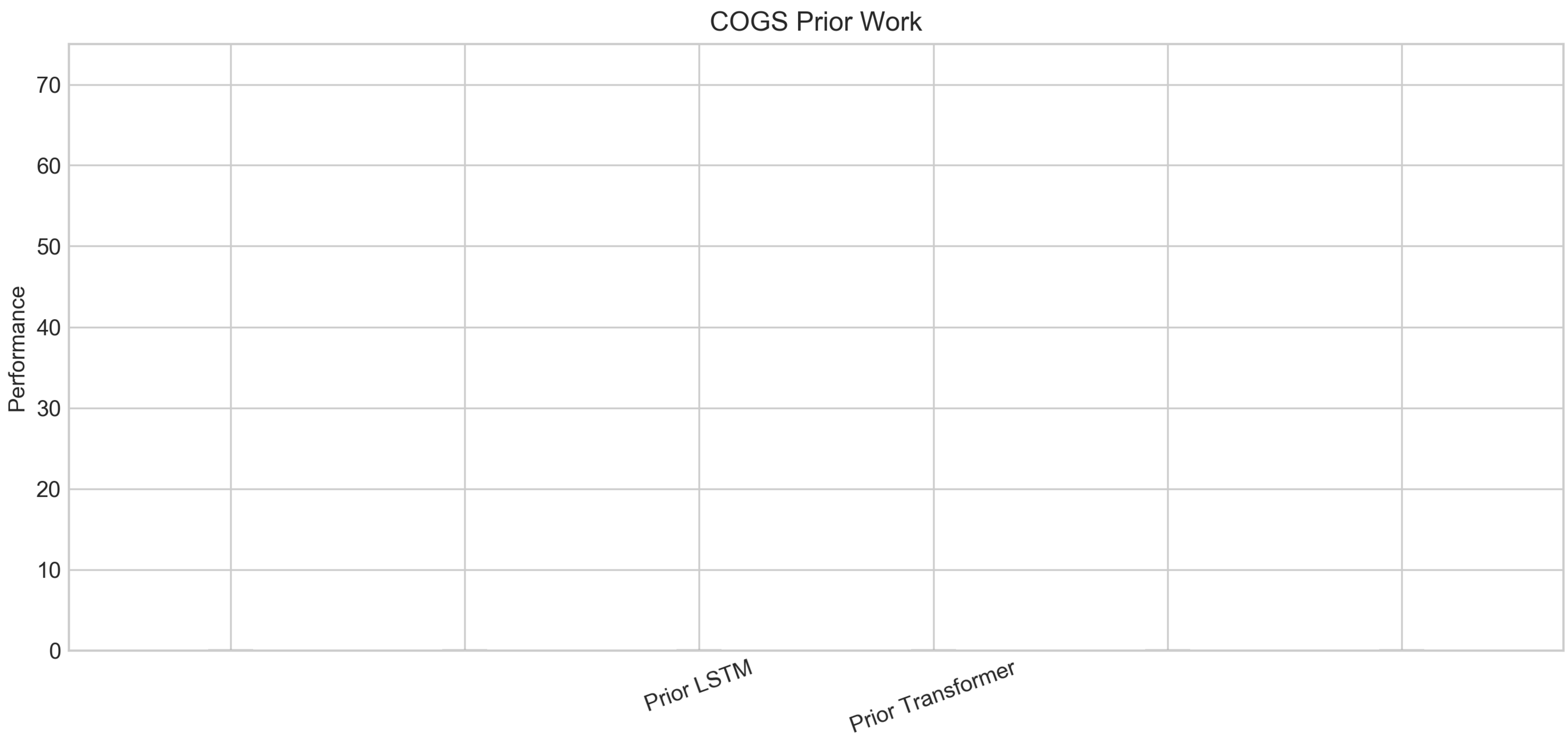
COGS (Kim and Linzen, 2020)

`The sailor dusted a boy .` ⟶ `* sailor ( x _ 1 ) ; dust . agent ( x _ 2 , x _ 1 ) AND dust . theme ( x _ 2 , x _ 4 ) AND boy ( x _ 4 )`

Novel Combination of Familiar Primitives and Grammatical Roles	Subject ⟶ Object	`A hedgehog ate the cake.`	`The baby liked the hedgehog.`
Deeper Recursion	Sentential complements	`Emma said that Noah knew that the cat danced.`	`Emma said that Noah knew that Lucas saw that the cat danced.`

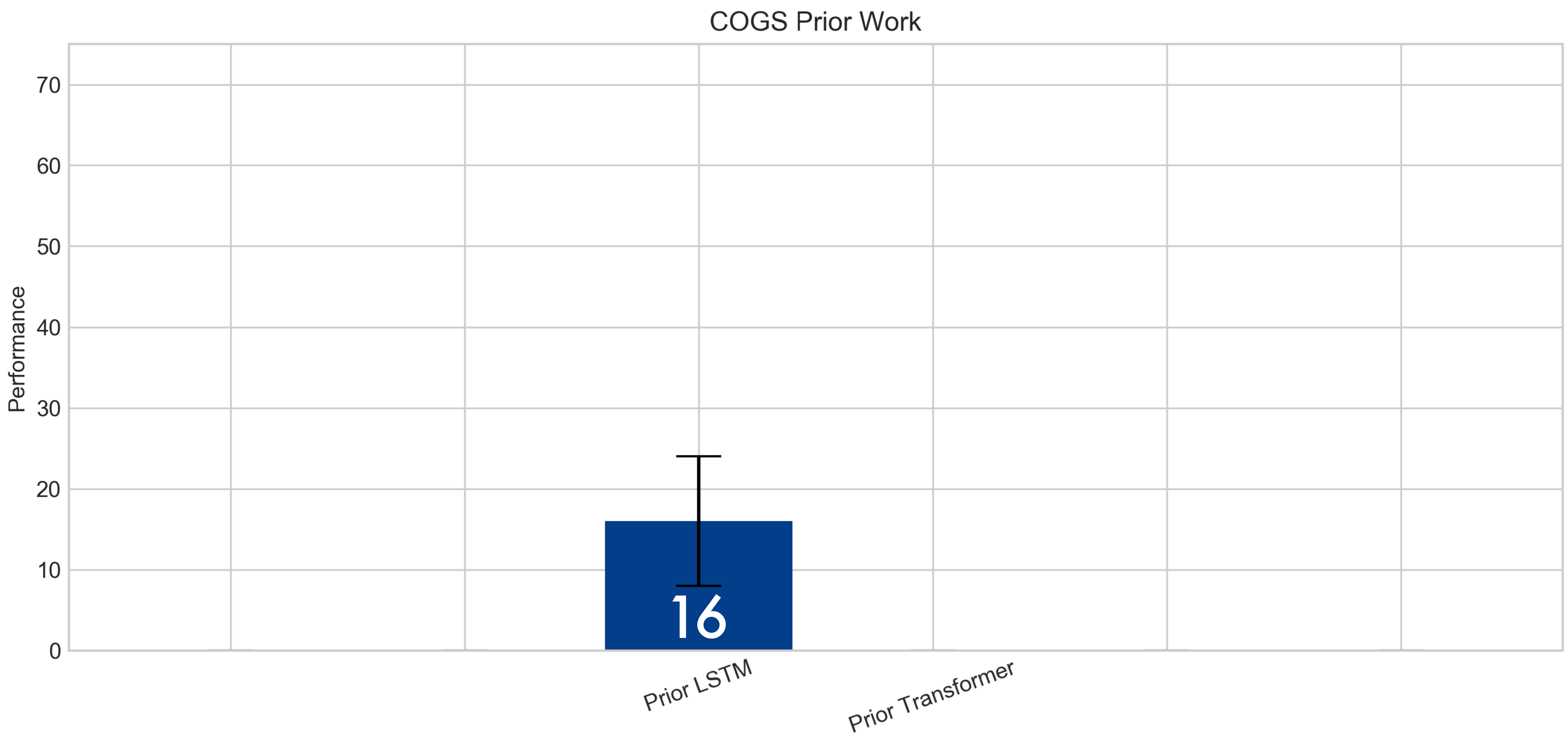
Explicitly testing compositionality

COGS (Kim and Linzen, 2020)



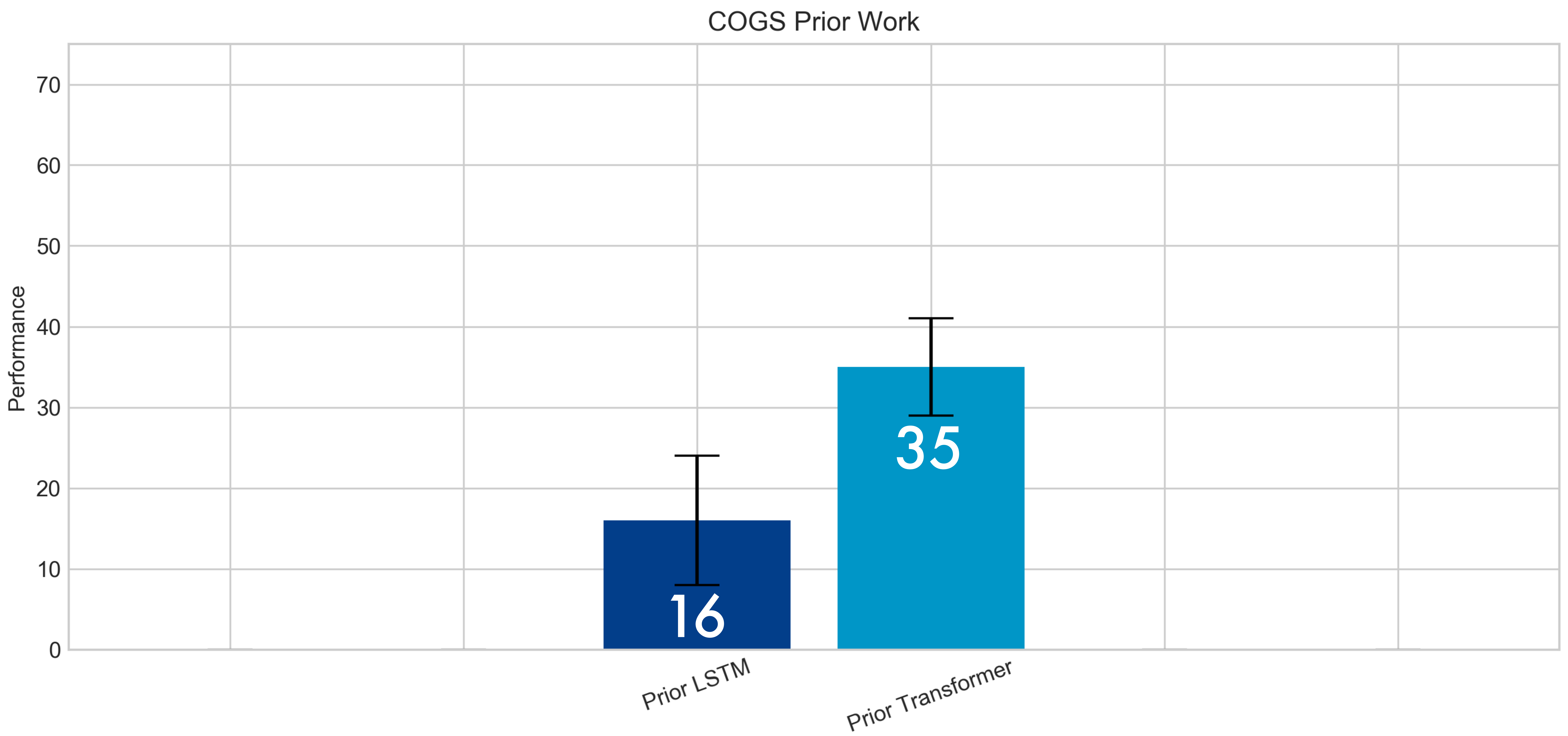
Explicitly testing compositionality

COGS (Kim and Linzen, 2020)



Explicitly testing compositionality

COGS (Kim and Linzen, 2020)



is it surprising that these tasks are hard?

is it surprising that this is hard?

Underspecification & Supervised learning

is it surprising that this is hard?

Underspecification & Supervised learning

- Data under-specifies for the generalizations that produced it (Goodman, 1955)

is it surprising that this is hard?

Underspecification & Supervised learning

- Data under-specifies for the generalizations that produced it   (Goodman, 1955)

- Models are trained on these tasks using supervised learning.

is it surprising that this is hard?

Underspecification & Supervised learning

- Data under-specifies for the generalizations that produced it (Goodman, 1955)

- Models are trained on these tasks using supervised learning.
- Independent and Identically Distributed Assumption

is it surprising that this is hard?

Underspecification & Supervised learning

- Data under-specifies for the generalizations that produced it (Goodman, 1955)

- Models are trained on these tasks using supervised learning.
- Independent and Identically Distributed Assumption
  - explicitly not met here.

is it surprising that this is hard?

Underspecification & Supervised learning

- Data under-specifies for the generalizations that produced it (Goodman, 1955)

- Models are trained on these tasks using supervised learning.
- Independent and Identically Distributed Assumption
  - explicitly not met here.

- **Underspecified data + Supervised learning may fail to consistently extract robust strategies**

Improving Generalization

DG-MAML (Li et al., 2018)

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

no-embedding | dataset | lots of embedding

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

I like the cat.

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

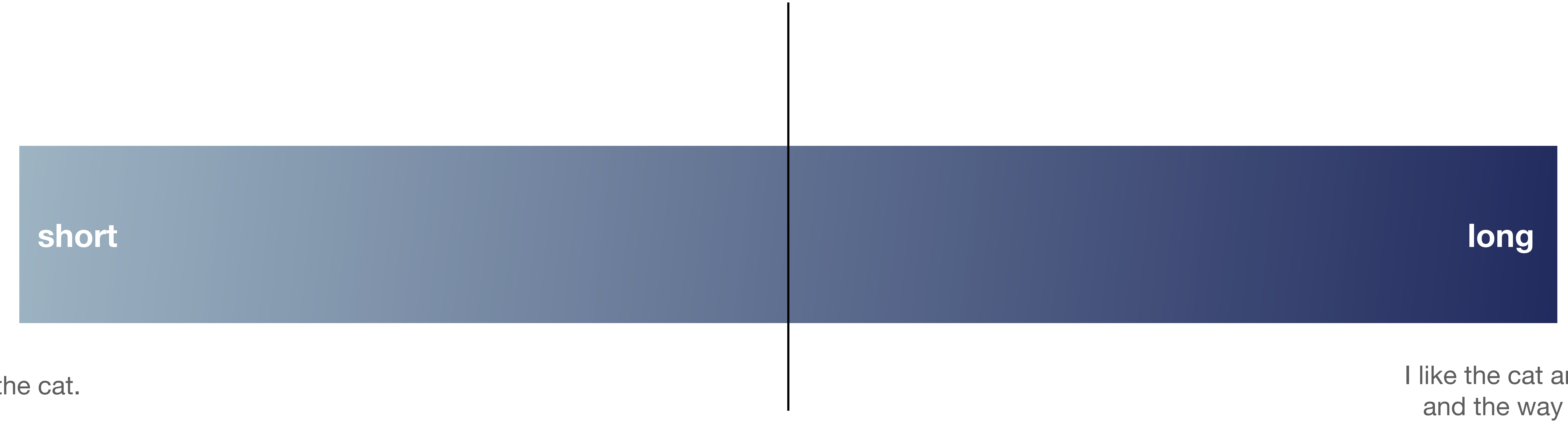
I like the cat.

I like the cat and the dog and the way that they seem to be friends.

Improving Generalization

DG-MAML (Li et al., 2018)

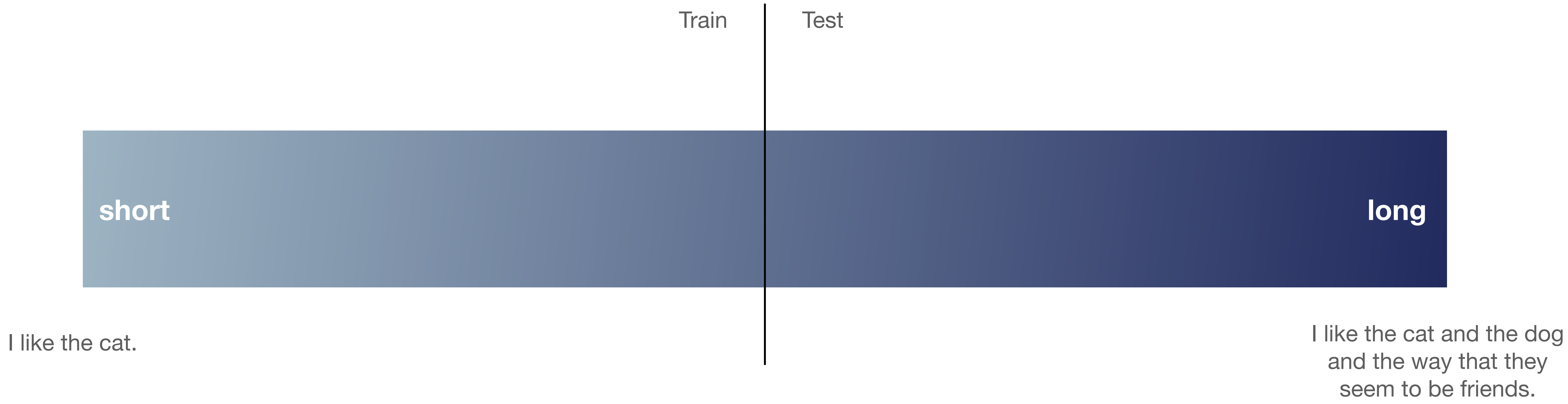
- if we know we’re going to be tested on something different, let’s train for that



Improving Generalization

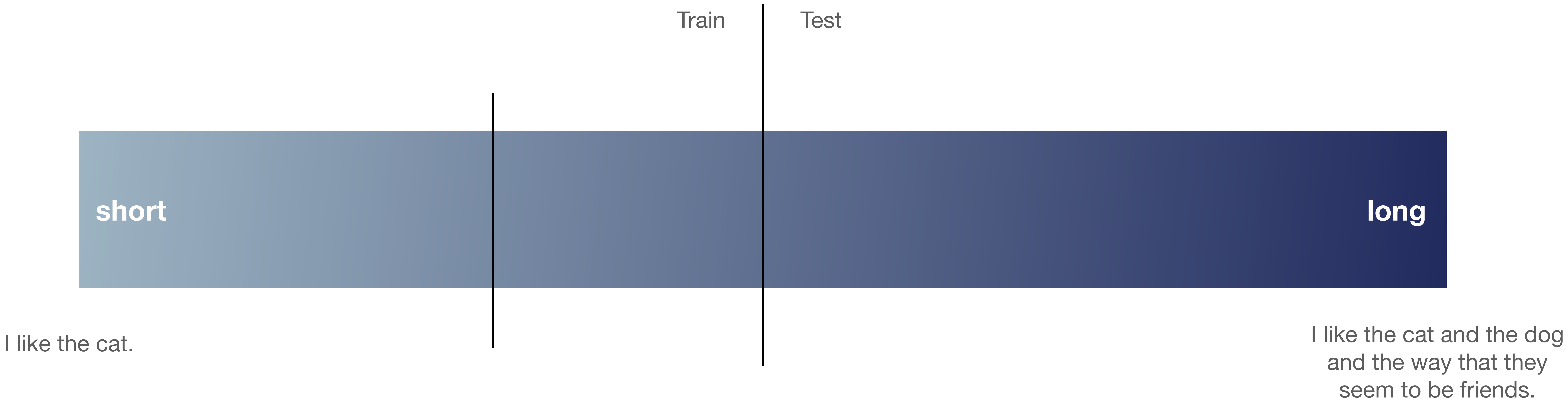
DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that



Improving Generalization
DG-MAML (Li et al., 2018)

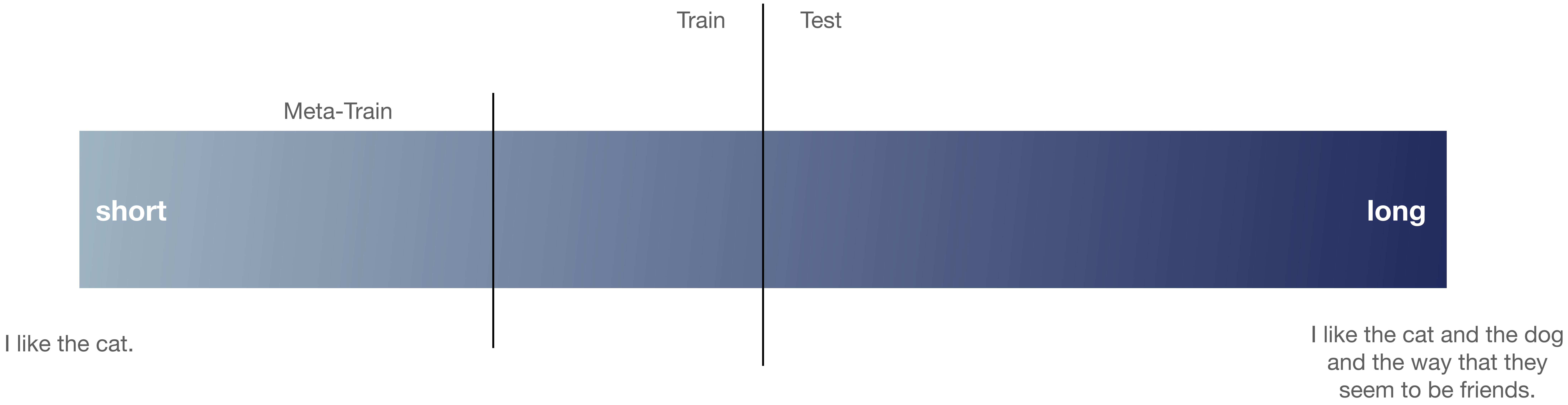
- if we know we're going to be tested on something different, let's train for that



Improving Generalization

DG-MAML (Li et al., 2018)

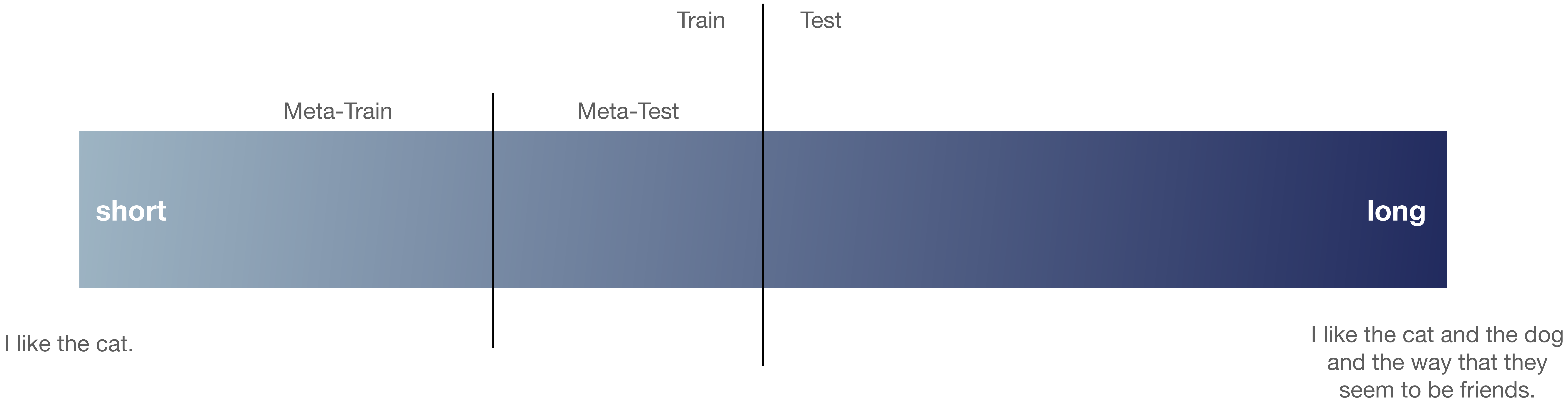
- if we know we're going to be tested on something different, let's train for that



Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that



Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$\mathcal{T}$

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

τ

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$\mathcal{B}_t$ (sampled from meta-train)

τ

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$$\mathcal{T} \begin{cases} \mathcal{B}_t & \text{(sampled from meta-train)} \\ \mathcal{B}_g & \text{(sampled from meta-test)} \end{cases}$$

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$\mathcal{T}$

$\mathcal{B}_t$ (sampled from meta-train)

$\mathcal{B}_g$ (sampled from meta-test)

$$\mathcal{L}_{\mathcal{B}_t}(\theta)$$

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$\mathcal{T}$

$\mathcal{B}_t$ (sampled from meta-train)

$\mathcal{B}_g$ (sampled from meta-test)

$$\mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta)$$

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$\mathcal{T}$ → $\mathcal{B}_t$ (sampled from meta-train)

$\mathcal{T}$ → $\mathcal{B}_g$ (sampled from meta-test)

$$\mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\mathcal{L}_{\mathcal{B}_g}(\theta')$$

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$$\mathcal{T} \begin{cases} \mathcal{B}_t & \text{(sampled from meta-train)} \\ \mathcal{B}_g & \text{(sampled from meta-test)} \end{cases}$$

$$\mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\mathcal{L}_{\mathcal{B}_g}(\theta')$$

$$\mathcal{L}_{\mathcal{B}_t}(\theta)$$

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$\mathcal{T}$ → $\mathcal{B}_t$ (sampled from meta-train)

$\mathcal{T}$ → $\mathcal{B}_g$ (sampled from meta-test)

$$\mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\mathcal{L}_{\mathcal{B}_g}(\theta')$$

$$\mathcal{L}_{\mathcal{B}_t}(\theta) + \mathcal{L}_{\mathcal{B}_g}(\theta')$$

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$$\mathcal{T} \begin{cases} \mathcal{B}_t & \text{(sampled from meta-train)} \\ \mathcal{B}_g & \text{(sampled from meta-test)} \end{cases}$$

$$\mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\mathcal{L}_{\mathcal{B}_g}(\theta')$$

$$\mathcal{L}_{\mathcal{T}}(\theta) = \mathcal{L}_{\mathcal{B}_t}(\theta) + \mathcal{L}_{\mathcal{B}_g}(\theta')$$

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$\mathcal{T}$

$\mathcal{B}_t$ (sampled from meta-train)

$\mathcal{B}_g$ (sampled from meta-test)

$$\mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\mathcal{L}_{\mathcal{B}_g}(\theta')$$

$$\mathcal{L}_{\mathcal{T}}(\theta) = \mathcal{L}_{\mathcal{B}_t}(\theta) + \mathcal{L}_{\mathcal{B}_g}(\theta')$$

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$\mathcal{T}$

$\mathcal{B}_t$ (sampled from meta-train)

$\mathcal{B}_g$ (sampled from meta-test)

$$\mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\mathcal{L}_{\mathcal{B}_g}(\theta')$$

$$\mathcal{L}_{\mathcal{T}}(\theta) = \mathcal{L}_{\mathcal{B}_t}(\theta) + \mathcal{L}_{\mathcal{B}_g}(\theta')$$

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$\mathcal{T}$

$\mathcal{B}_t$ (sampled from meta-train)

$\mathcal{B}_g$ (sampled from meta-test)

$$\mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\mathcal{L}_{\mathcal{B}_g}(\theta')$$

$$\mathcal{L}_{\mathcal{T}}(\theta) = \mathcal{L}_{\mathcal{B}_t}(\theta) + \mathcal{L}_{\mathcal{B}_g}(\theta')$$

$$\mathcal{L}_{\mathcal{T}}(\theta)$$

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$$\mathcal{T} \quad \begin{cases} \mathcal{B}_t & \text{(sampled from meta-train)} \\ \mathcal{B}_g & \text{(sampled from meta-test)} \end{cases}$$

$$\mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\mathcal{L}_{\mathcal{B}_g}(\theta')$$

$$\mathcal{L}_{\mathcal{T}}(\theta) = \mathcal{L}_{\mathcal{B}_t}(\theta) + \mathcal{L}_{\mathcal{B}_g}(\theta')$$

$$\theta^* \leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{\mathcal{T}}(\theta)$$

Improving Generalization

DG-MAML (Li et al., 2018)

- if we know we're going to be tested on something different, let's train for that

$\mathcal{T}$

$\mathcal{B}_t$ (sampled from meta-train)

$\mathcal{B}_g$ (sampled from meta-test)

$$\mathcal{L}_{\mathcal{B}_t}(\theta)$$

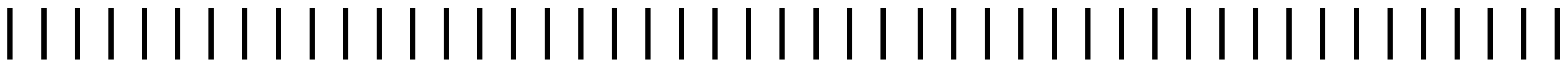
$$\theta' \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{B}_t}(\theta)$$

$$\mathcal{L}_{\mathcal{B}_g}(\theta')$$

$$\boxed{\begin{aligned} \mathcal{L}_{\mathcal{T}}(\theta) &= \mathcal{L}_{\mathcal{B}_t}(\theta) + \mathcal{L}_{\mathcal{B}_g}(\theta') \\ \theta^* &\leftarrow \theta - \beta \nabla_\theta \mathcal{L}_{\mathcal{T}}(\theta) \end{aligned}}$$

Improving Generalization

DG-MAML



Possible
Strategies

Improving Generalization

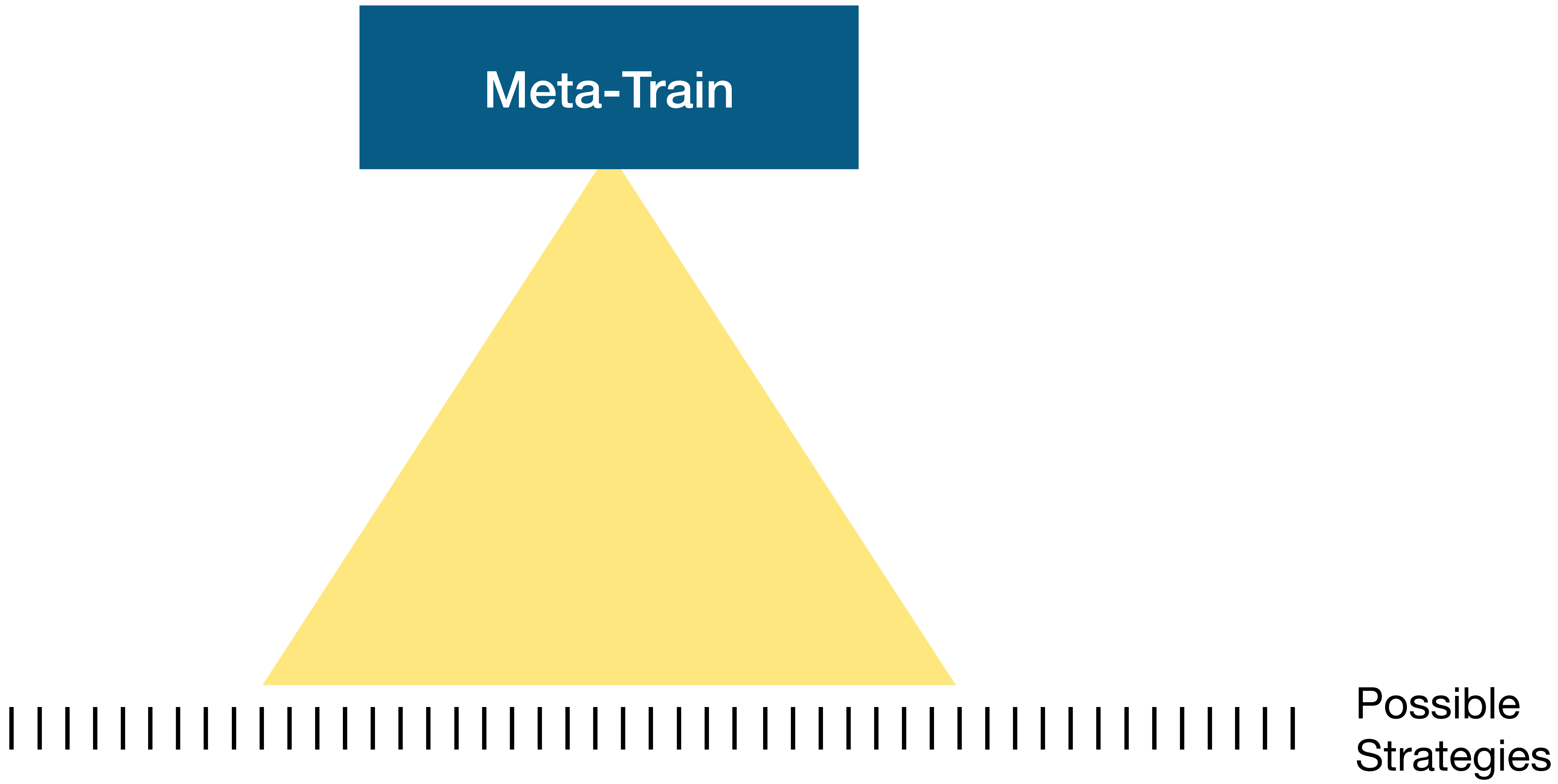
DG-MAML

Meta-Train

Possible
Strategies

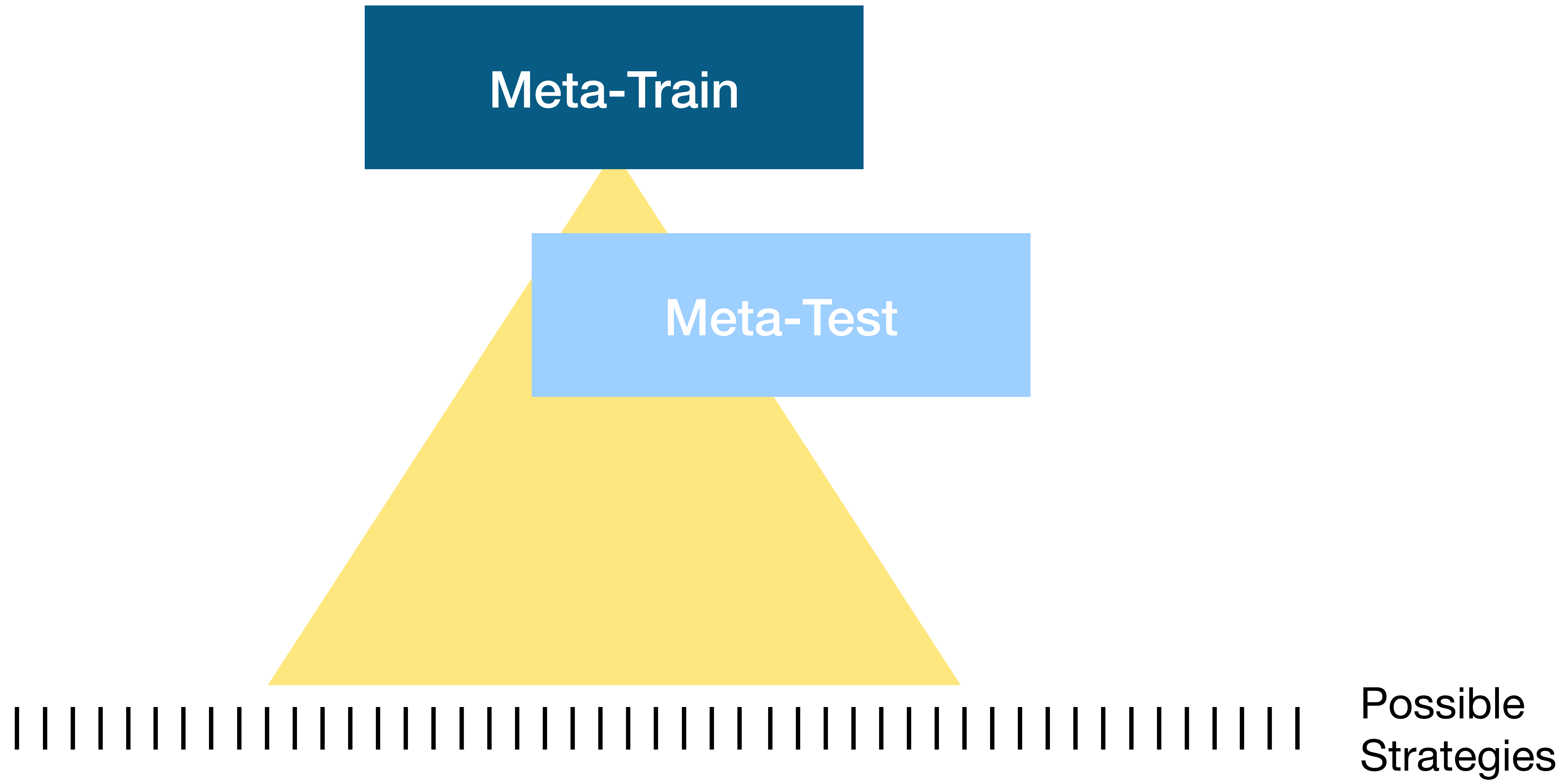
Improving Generalization

DG-MAML



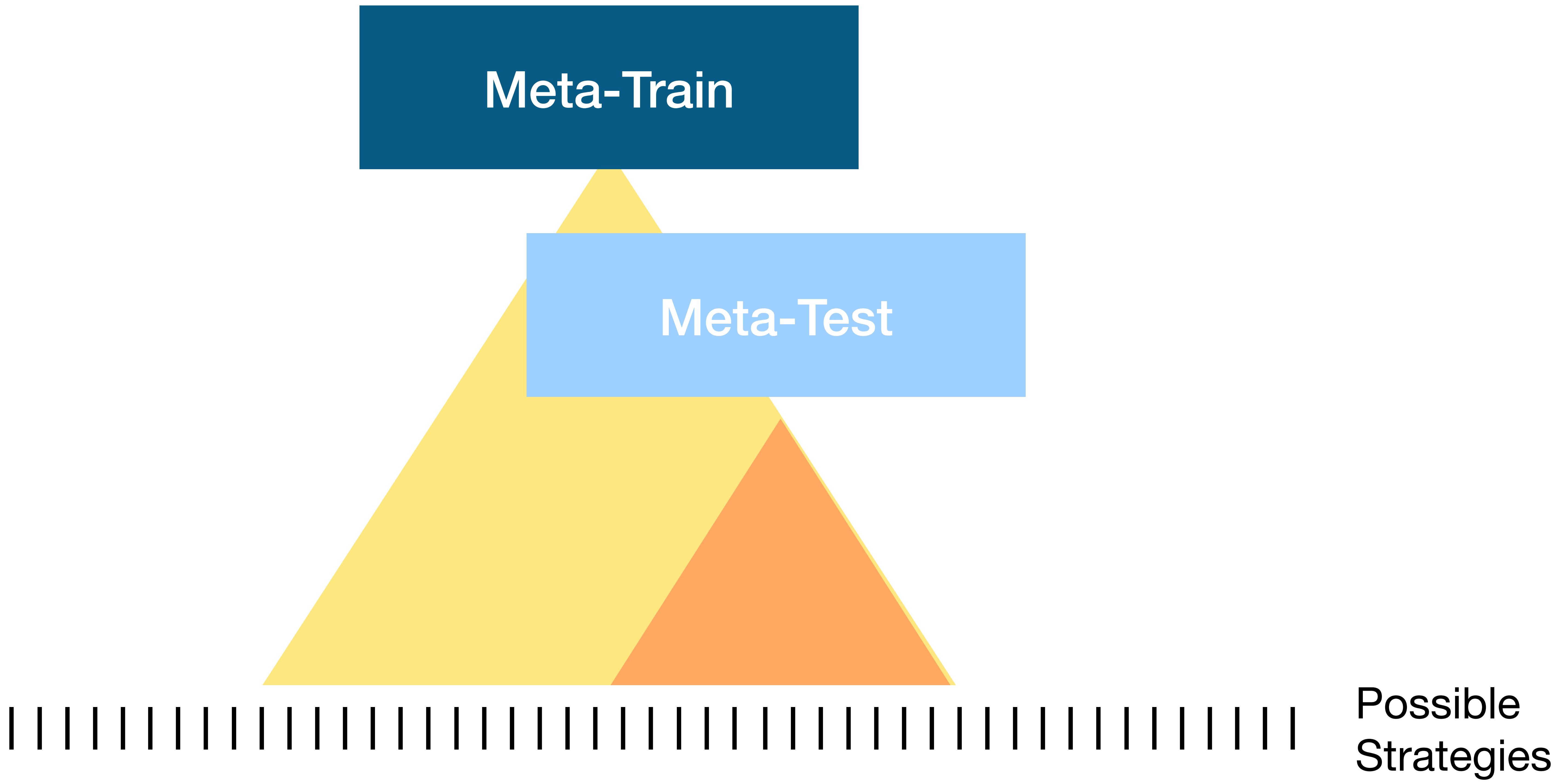
Improving Generalization

DG-MAML



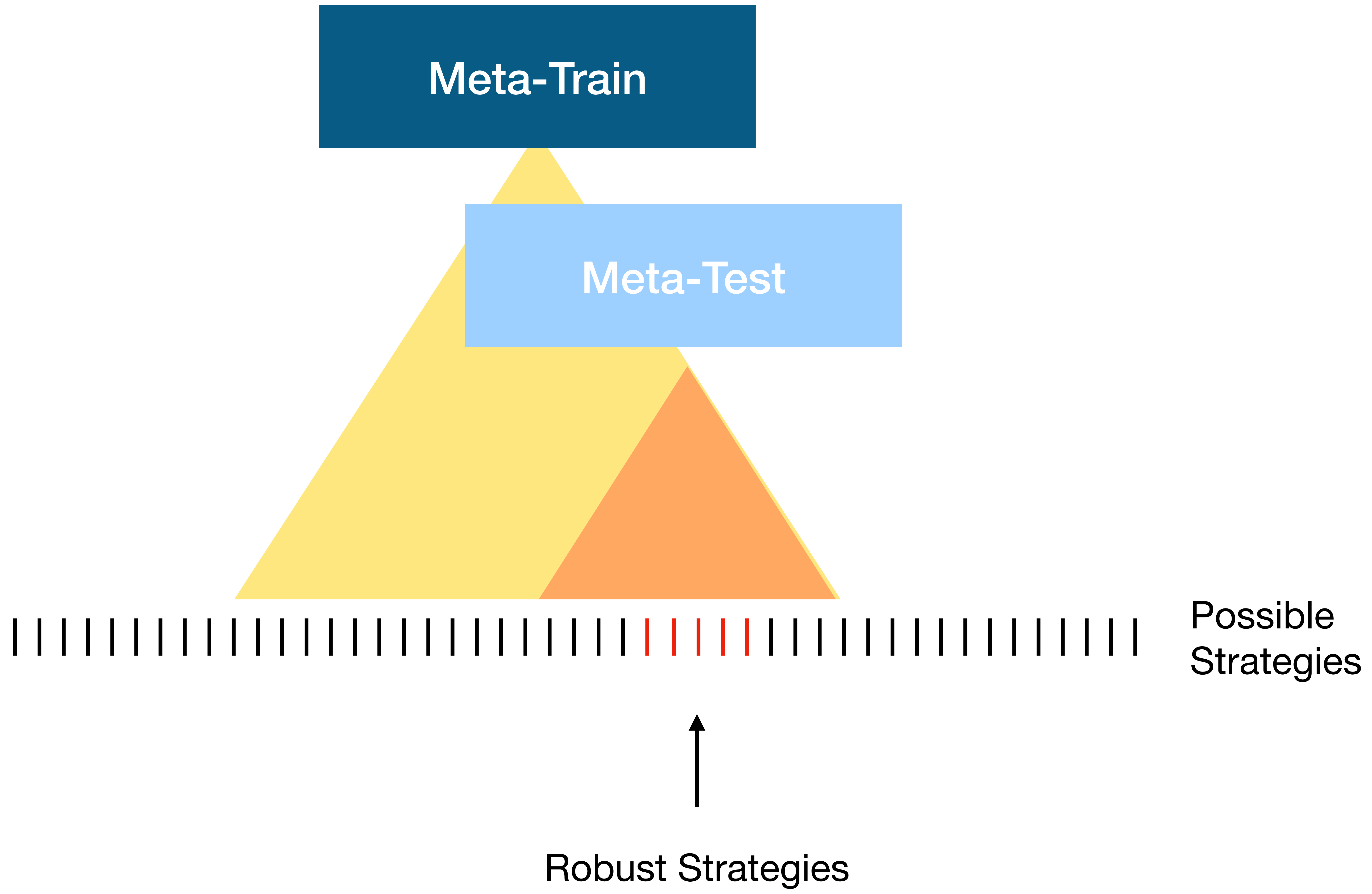
Improving Generalization

DG-MAML



Improving Generalization

DG-MAML



Improving Generalization

Prior Knowledge

Improving Generalization

Prior Knowledge

- But this requires prior knowledge

Improving Generalization

Prior Knowledge

- But this requires prior knowledge
  - not of the test distribution, but of the family of distributions from which the test distribution will be drawn

how can we help to resolve this underspecification more domain-generally?

Domain-General Bias

DG-MAML

Domain-General Bias

DG-MAML

- DG-MAML presents a way to introduce a bias during training

Domain-General Bias

DG-MAML

- DG-MAML presents a way to introduce a bias during training

$$\mathcal{L}_{\mathcal{T}}(\theta) = \mathcal{L}_{\mathcal{B}_t}(\theta) + \mathcal{L}_{\mathcal{B}_g}(\theta')$$

Domain-General Bias

DG-MAML

- DG-MAML presents a way to introduce a bias during training

$$\mathcal{L}_{\mathcal{T}}(\theta) = \mathcal{L}_{\mathcal{B}_t}(\theta) + \boxed{\mathcal{L}_{\mathcal{B}_g}(\theta')}$$

Domain-General Bias

DG-MAML

- DG-MAML presents a way to introduce a bias during training

$$\mathcal{L}_{\mathcal{T}}(\theta) = \mathcal{L}_{\mathcal{B}_t}(\theta) + \boxed{\mathcal{L}_{\mathcal{B}_g}(\theta')}$$

- Whatever we put in the meta-test batch constrains our update step on meta-train

Domain-General Bias

DG-MAML

- DG-MAML presents a way to introduce a bias during training

$$\mathcal{L}_{\mathcal{T}}(\theta) = \mathcal{L}_{\mathcal{B}_t}(\theta) + \boxed{\mathcal{L}_{\mathcal{B}_g}(\theta')}$$

- Whatever we put in the meta-test batch constrains our update step on meta-train

Let's use this to introduce a general bias rather than a task specific one

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train

Meta-Test

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train	Meta-Test
`The girl changed a sandwich by the bed.`	

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train	Meta-Test

```
The girl changed a sandwich by the bed.
```

```
The sailor dusted a boy.
```

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train	Uniform Sampling | Uni-MAML

```
The girl changed a sandwich by the bed.

The sailor dusted a boy.
```

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train	Uniform Sampling | Uni-MAML
`The girl changed a sandwich by the bed.`	`The penguin ate a donut.`
`The sailor dusted a boy.`	

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train	Uniform Sampling | Uni-MAML
`The girl changed a sandwich by the bed.`	`The penguin ate a donut.` `Amelia gave Emma a strawberry.`
`The sailor dusted a boy.`	

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train	Uniform Sampling | Uni-MAML
`The girl changed a sandwich by the bed.`	`The penguin ate a donut.` `Amelia gave Emma a strawberry.`
`The sailor dusted a boy.`	`A cat disintegrated a girl.`

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train	Uniform Sampling | Uni-MAML
`The girl changed a sandwich by the bed.`	`The penguin ate a donut.` `Amelia gave Emma a strawberry.`
`The sailor dusted a boy.`	`A cat disintegrated a girl.` `A visitor was posted a rose by a turtle.`

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train	Lev Distance | Lev-MAML

```
The girl changed a sandwich by the bed.

The sailor dusted a boy.
```

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train	Lev Distance | Lev-MAML
`The girl changed a sandwich by the bed.`	`The girl rolled a drink beside the table.`
`The sailor dusted a boy.`	

Domain-General Bias

Memorization

- let’s introduce a bias that impairs the model’s ability to memorize whole sentences

Meta-Train	Lev Distance | Lev-MAML
`The girl changed a sandwich by the bed.`	`The girl rolled a drink beside the table.` `The girl liked a dealer beside the table .`
`The sailor dusted a boy.`	

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train	Lev Distance | Lev-MAML
`The girl changed a sandwich by the bed.`	`The girl rolled a drink beside the table.` `The girl liked a dealer beside the table .`
`The sailor dusted a boy.`	`The sailor dusted a girl.`

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train	Lev Distance | Lev-MAML
`The girl changed a sandwich by the bed.`	`The girl rolled a drink beside the table.` `The girl liked a dealer beside the table .`
`The sailor dusted a boy.`	`The sailor dusted a girl.` `The girl dusted a boy.`

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train	Convolutional String Kernel | Str-MAML
`The girl changed a sandwich by the bed.`	`The girl rolled a drink beside the table.` `The girl liked a dealer beside the table .`
`The sailor dusted a boy.`	`The sailor dusted a girl.` `The girl dusted a boy.`

Domain-General Bias

Memorization

- let's introduce a bias that impairs the model's ability to memorize whole sentences

Meta-Train	Partial Tree Kernel | Tree-MAML
`The girl changed a sandwich by the bed.`	`Mateo dusted a boy .` `dust`
`The sailor dusted a boy.`	`A sandwich changed.` `A block was changed by the girl.`

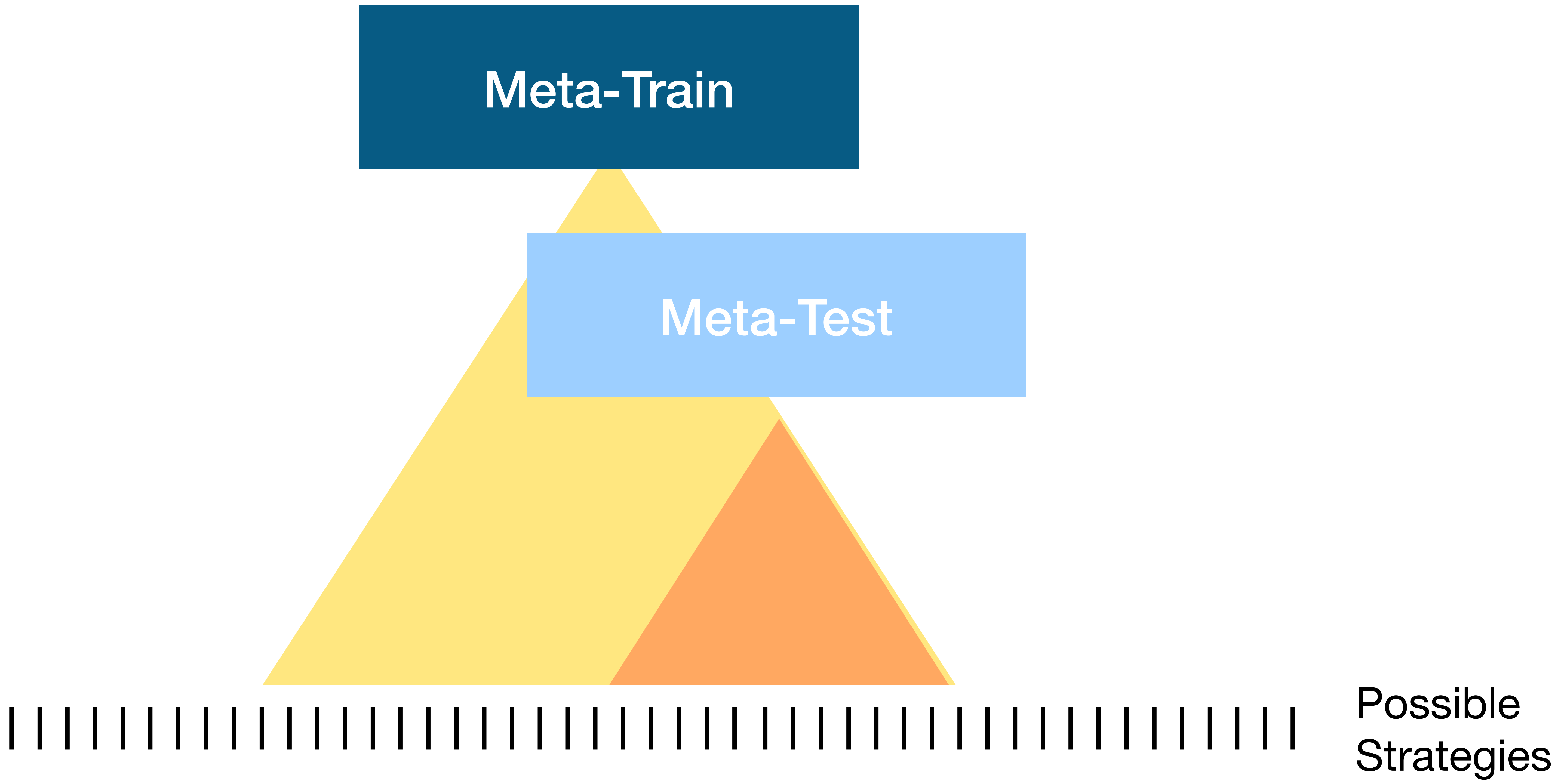
Does inhibiting models' memory improve generalization?

Experiments

COGS (& SCAN)

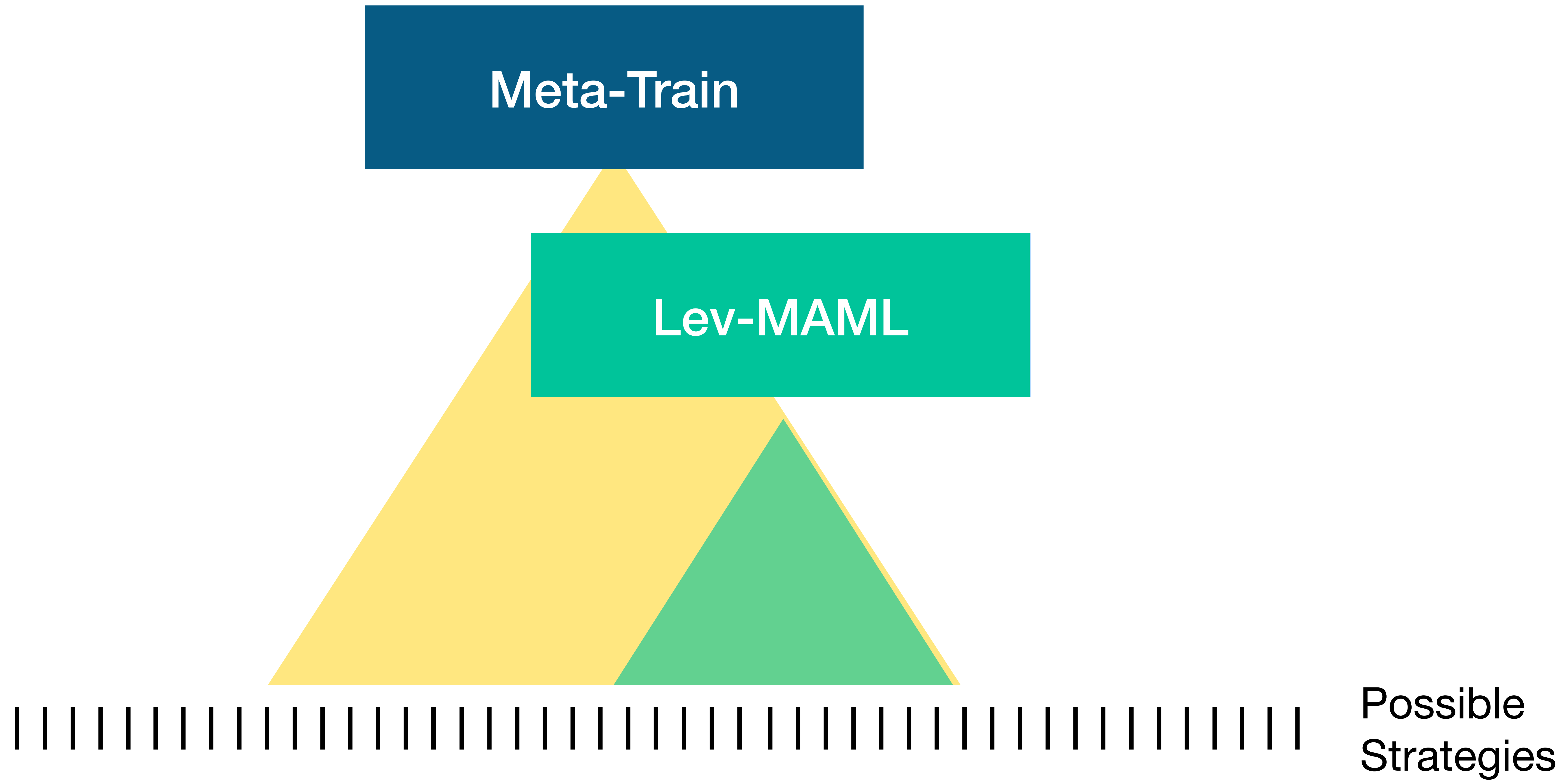
Experiments

COGS (& SCAN)



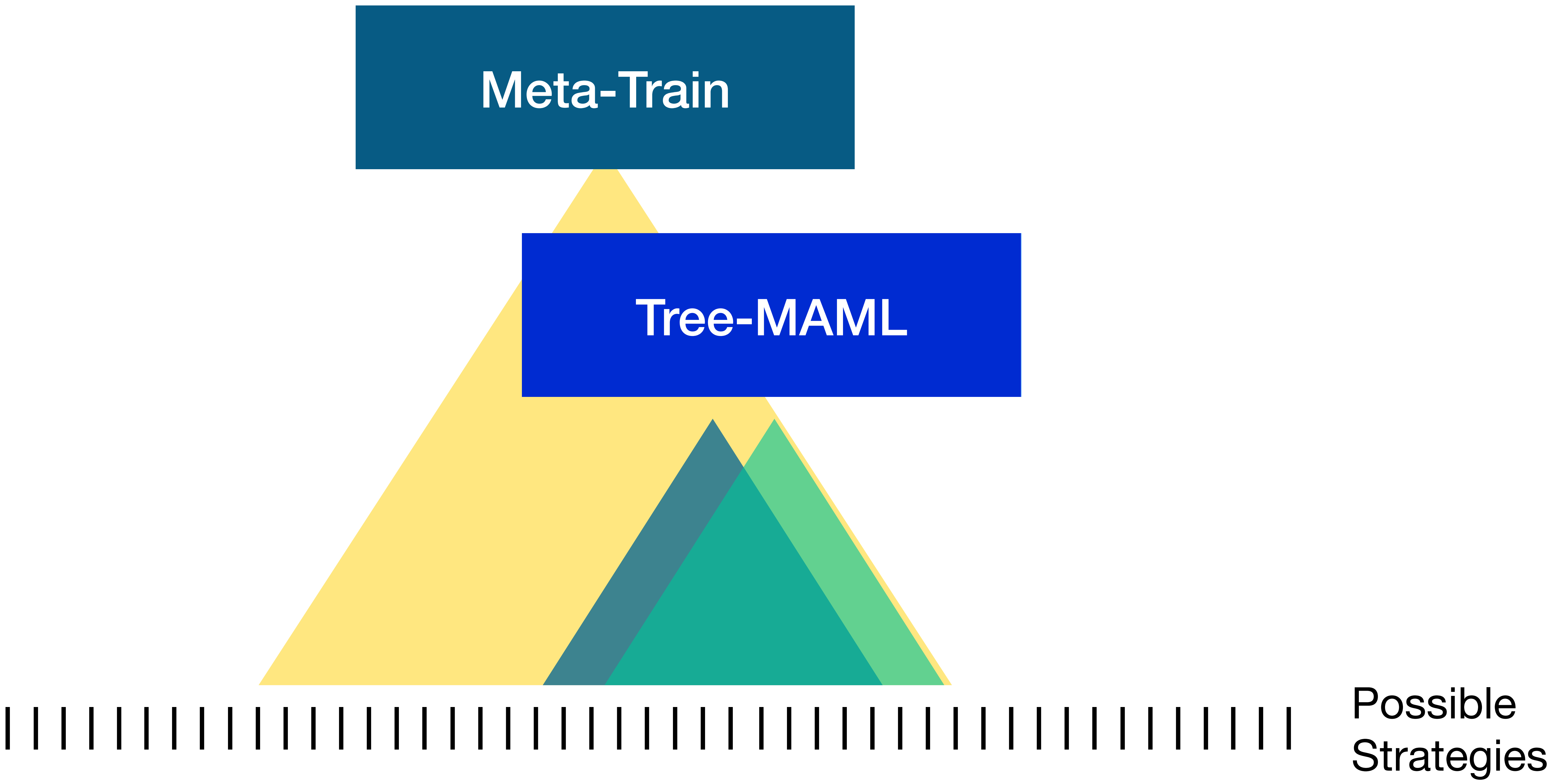
Experiments

COGS (& SCAN)



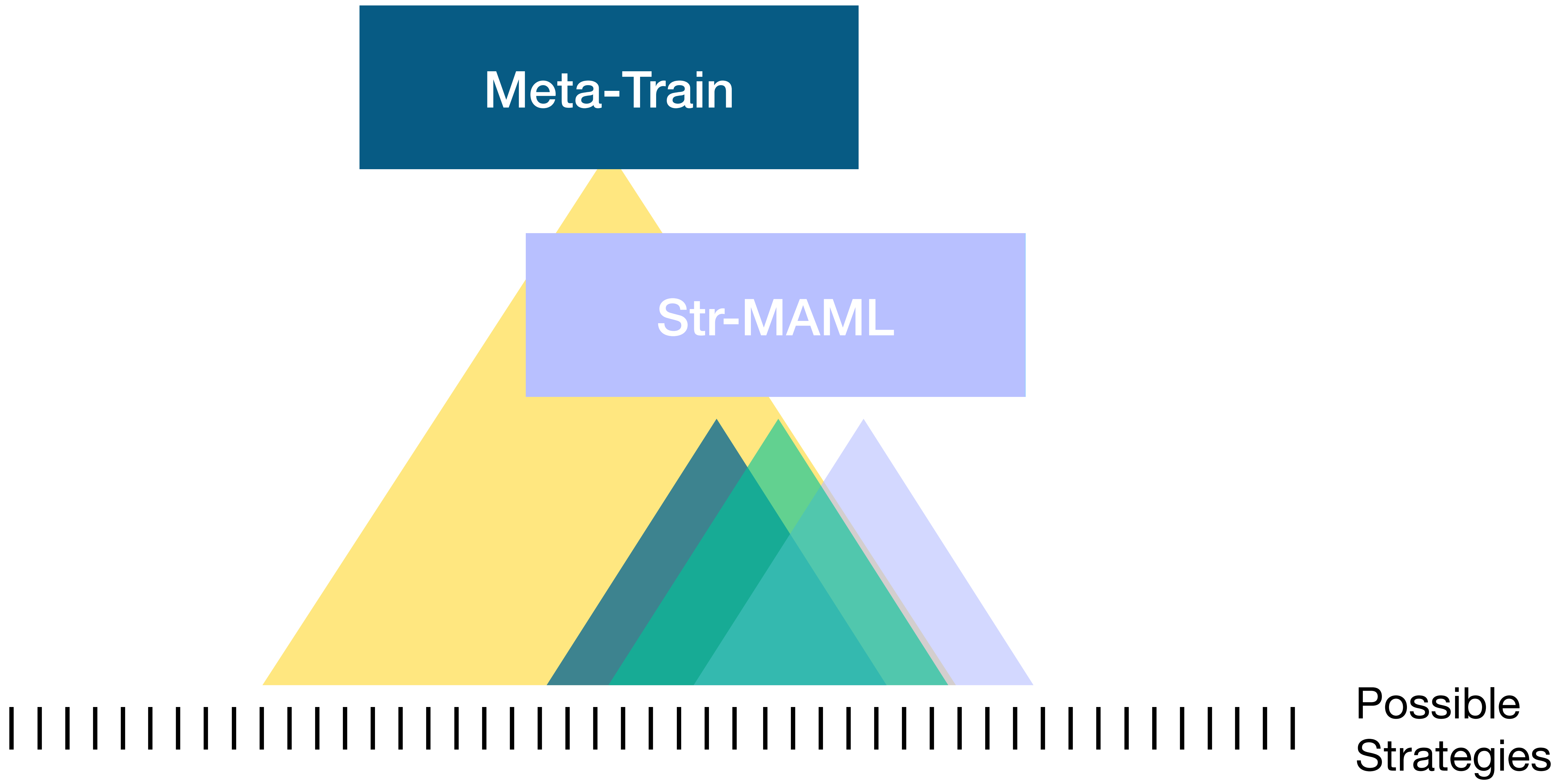
Experiments

COGS (& SCAN)



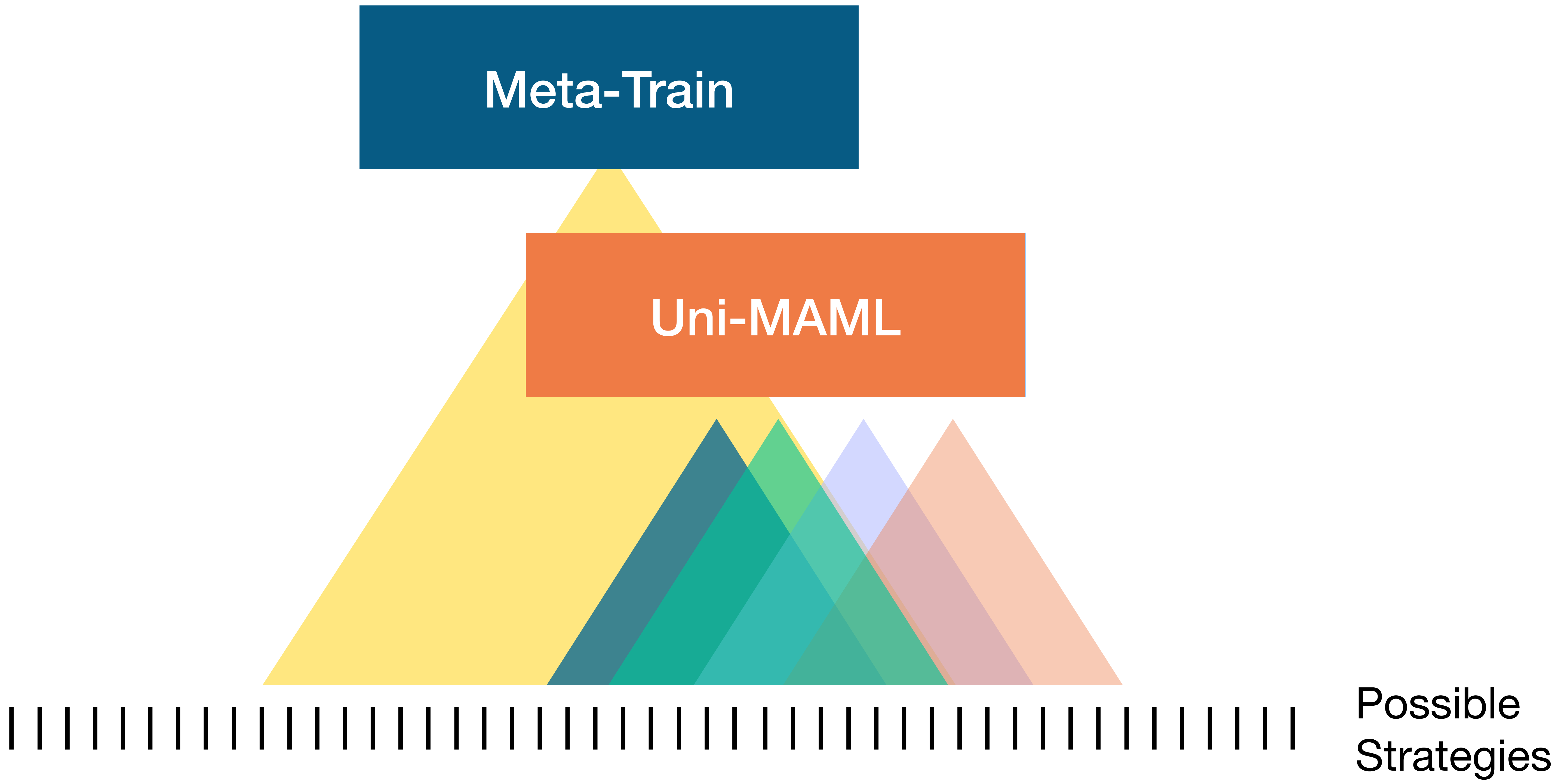
Experiments

COGS (& SCAN)



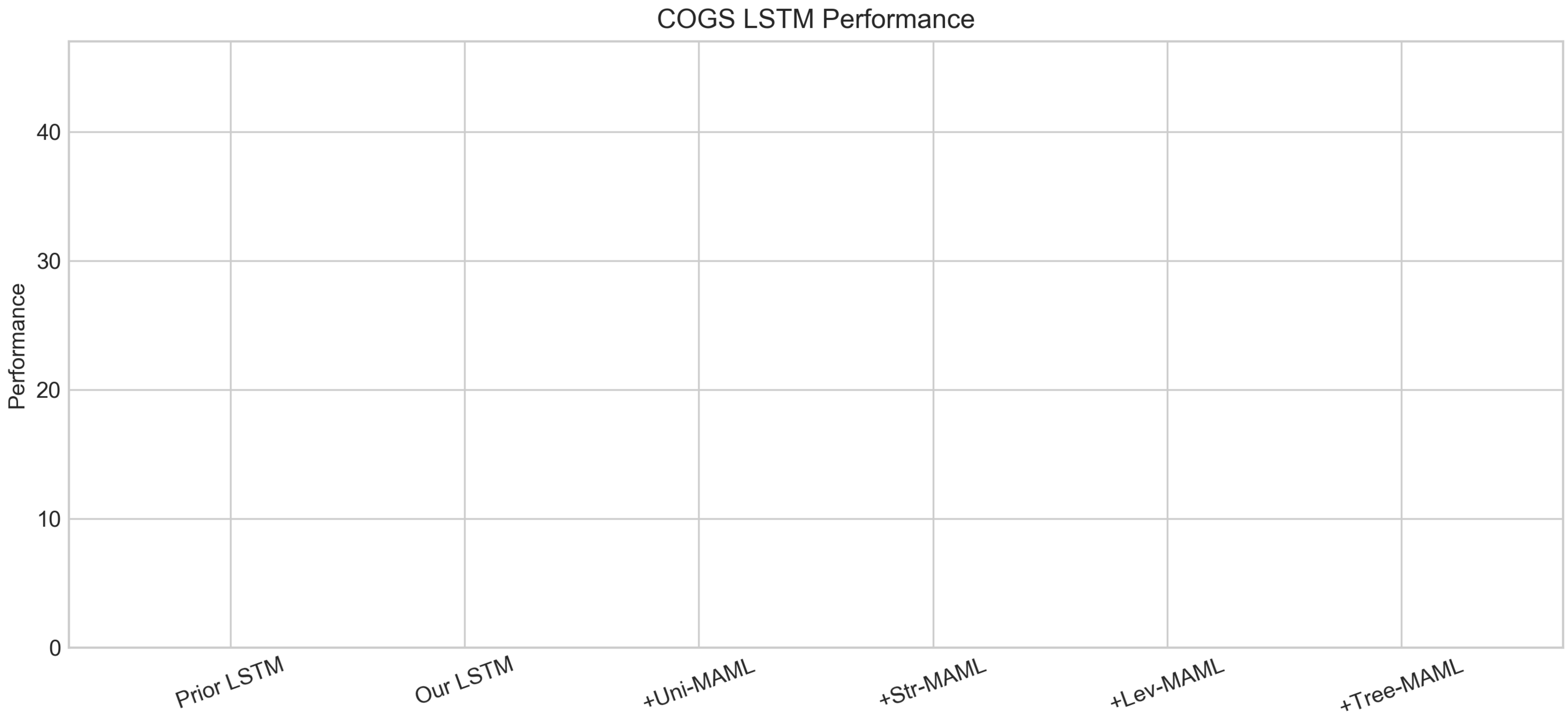
Experiments

COGS (& SCAN)

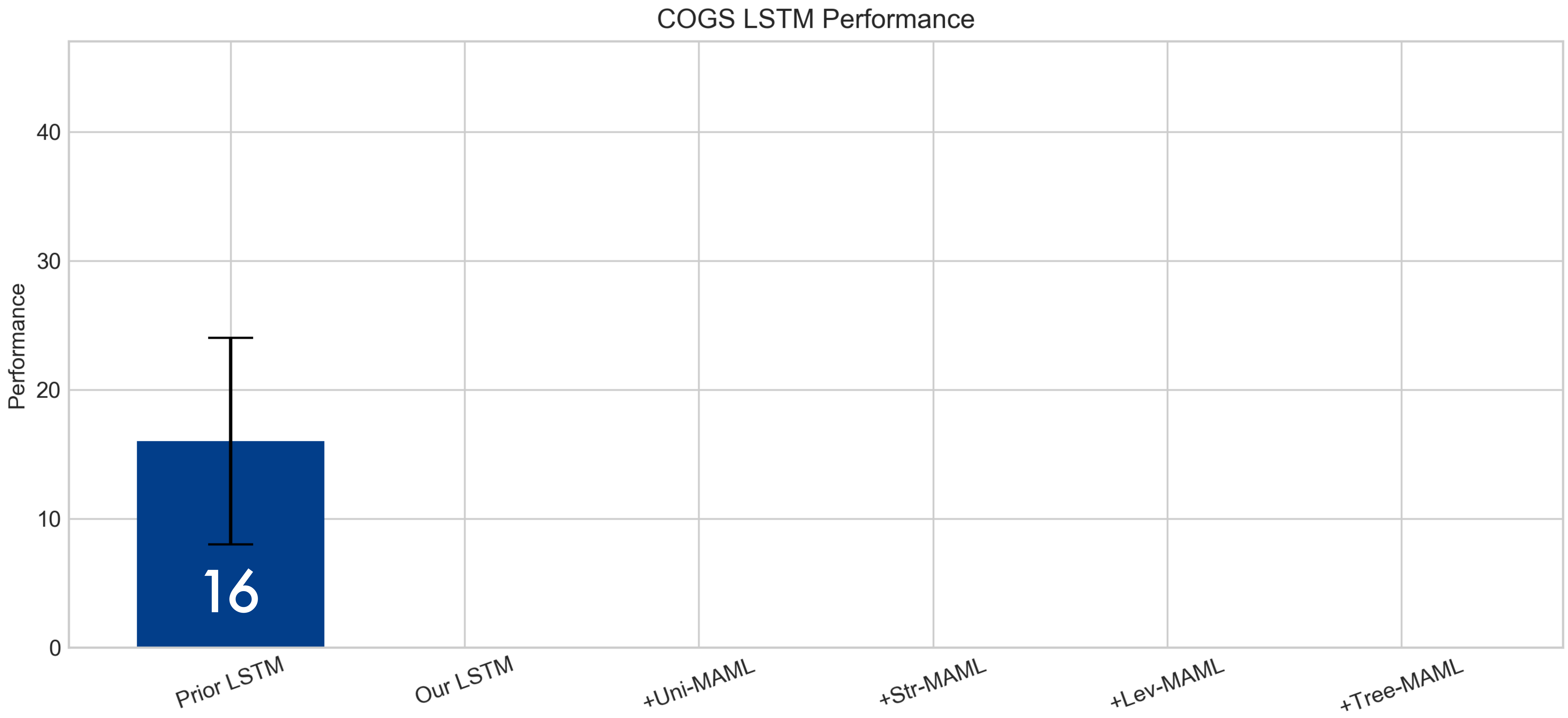


Experiments

COGS Dataset

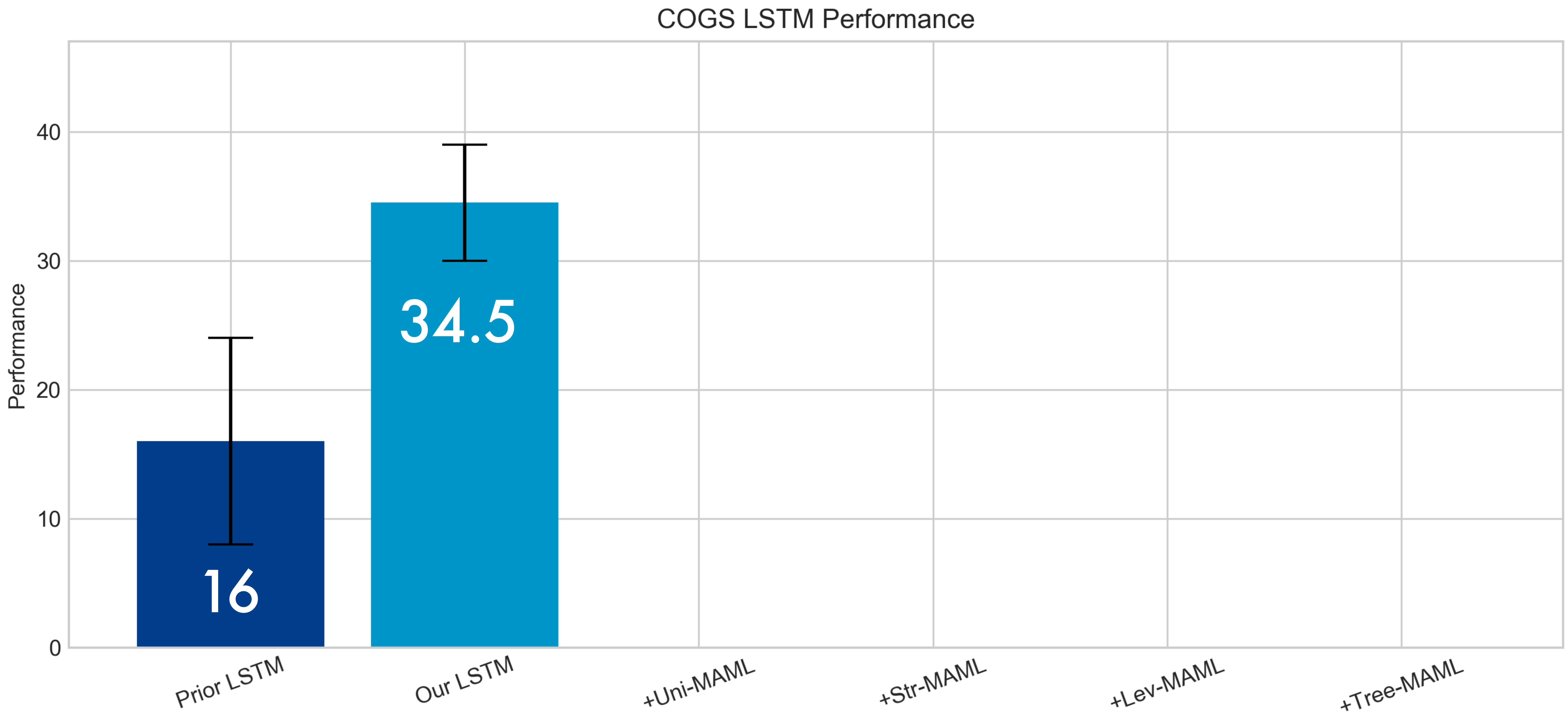


Experiments
COGS Dataset



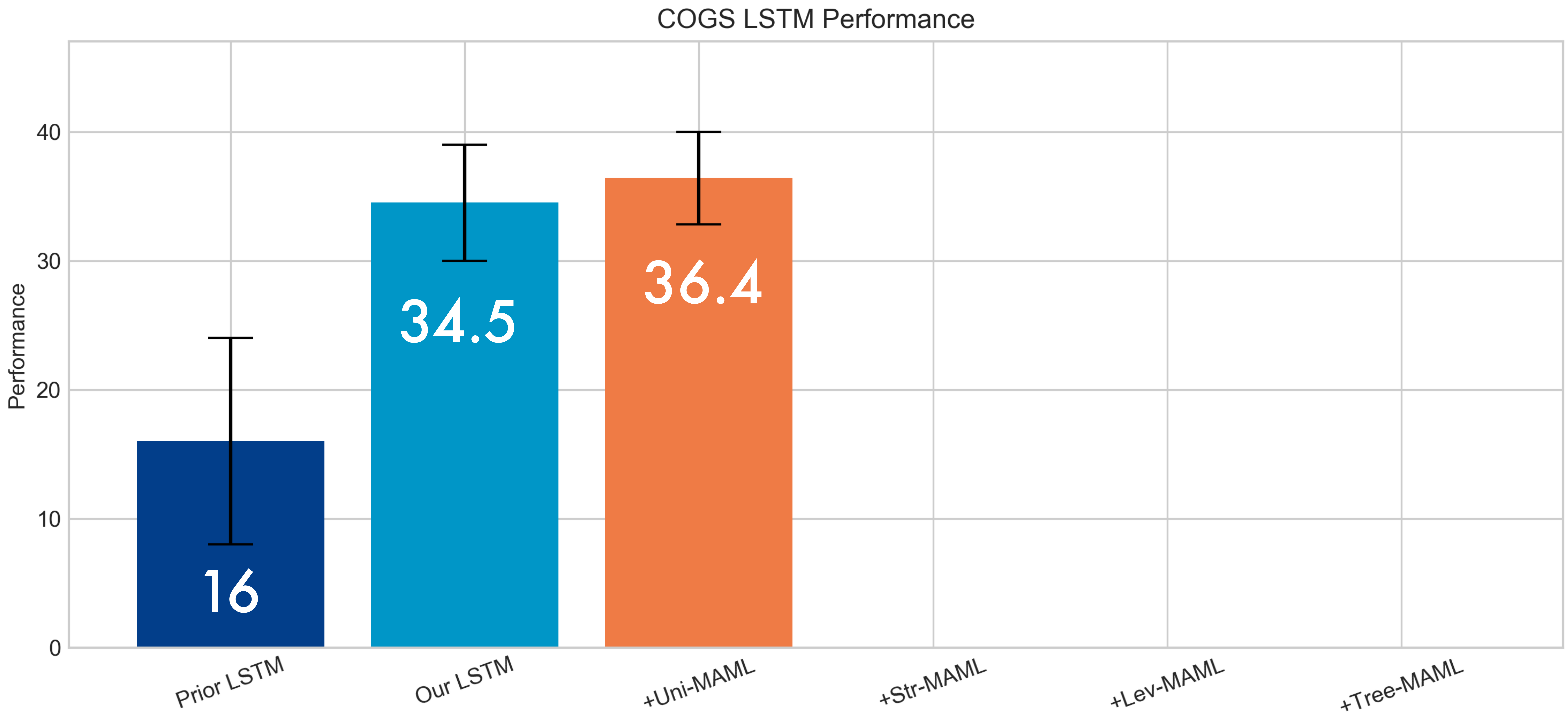
Experiments

COGS Dataset



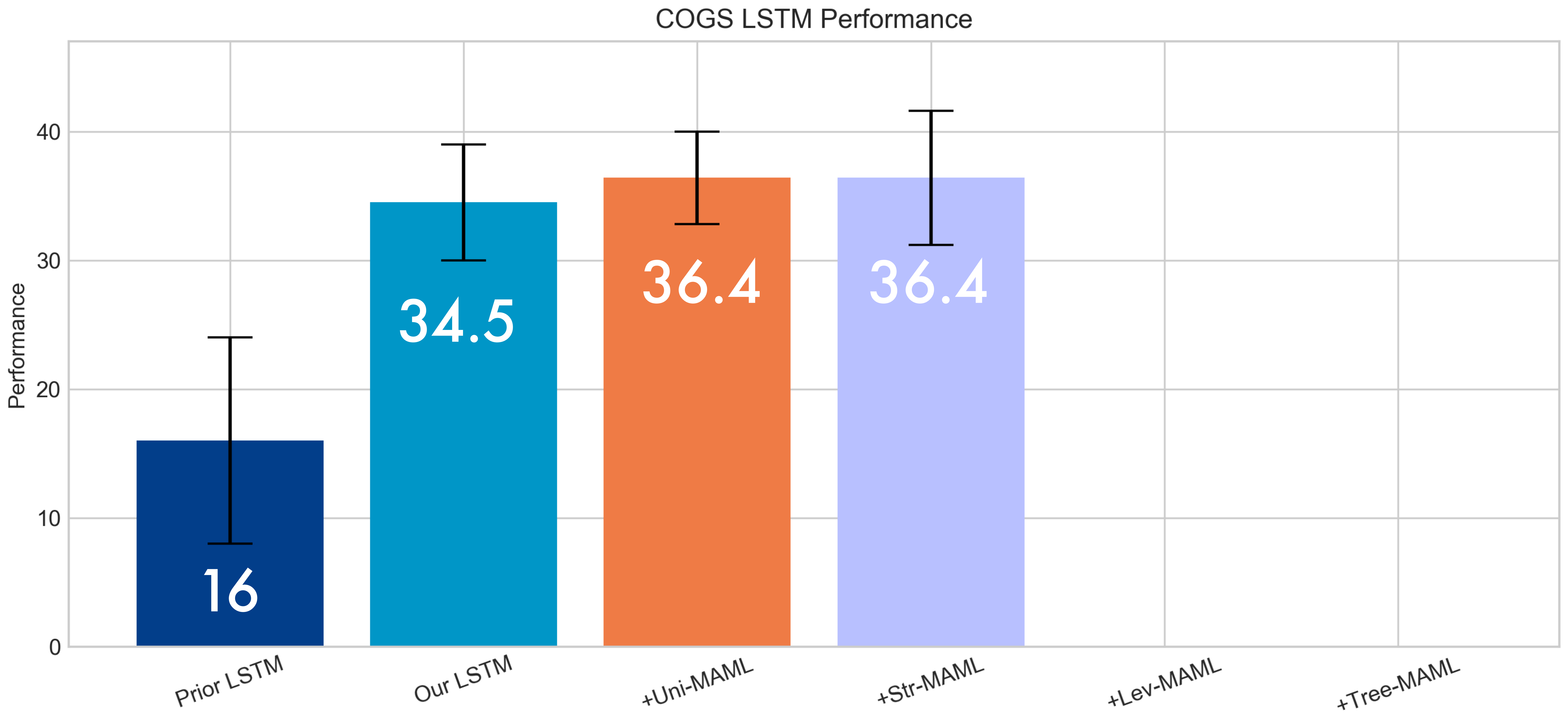
Experiments

COGS Dataset



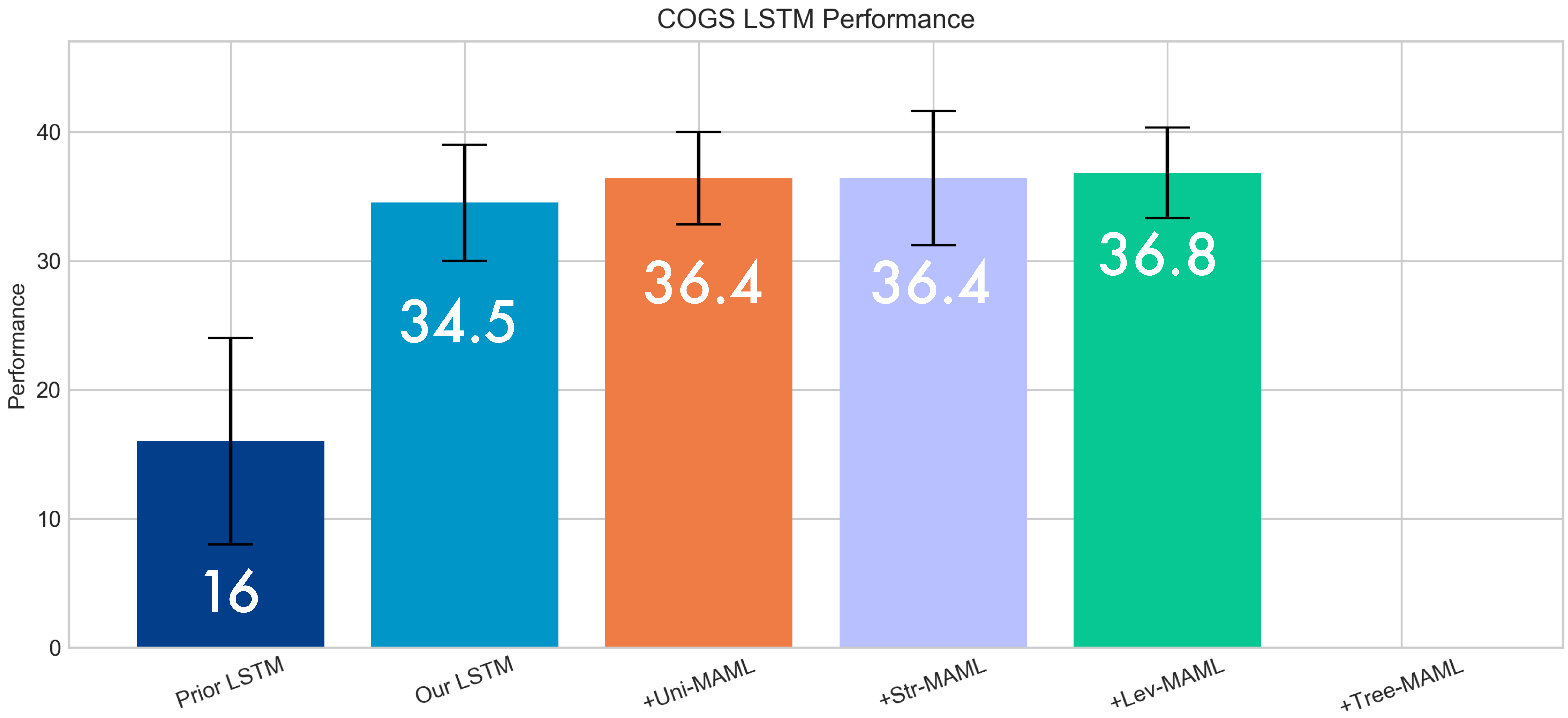
Experiments

COGS Dataset



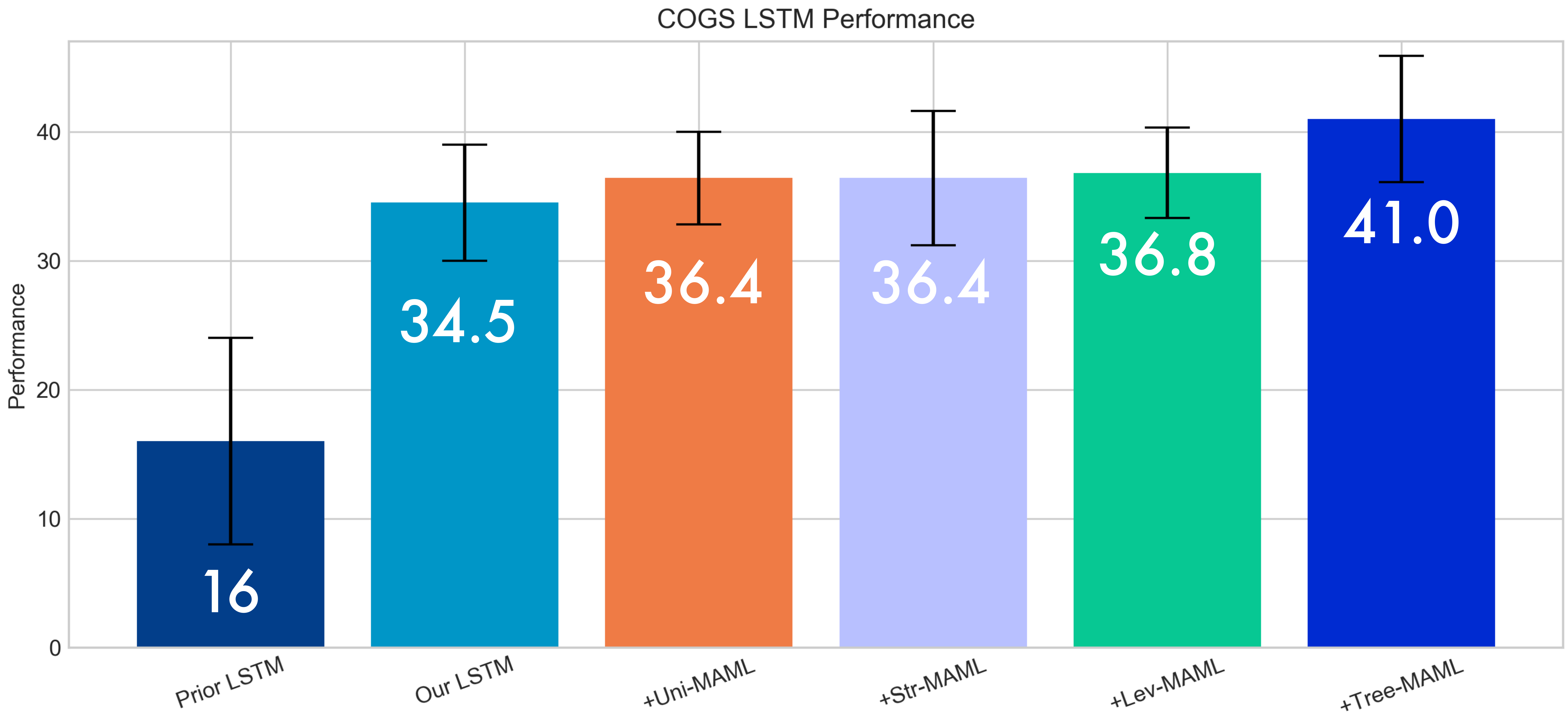
Experiments

COGS Dataset



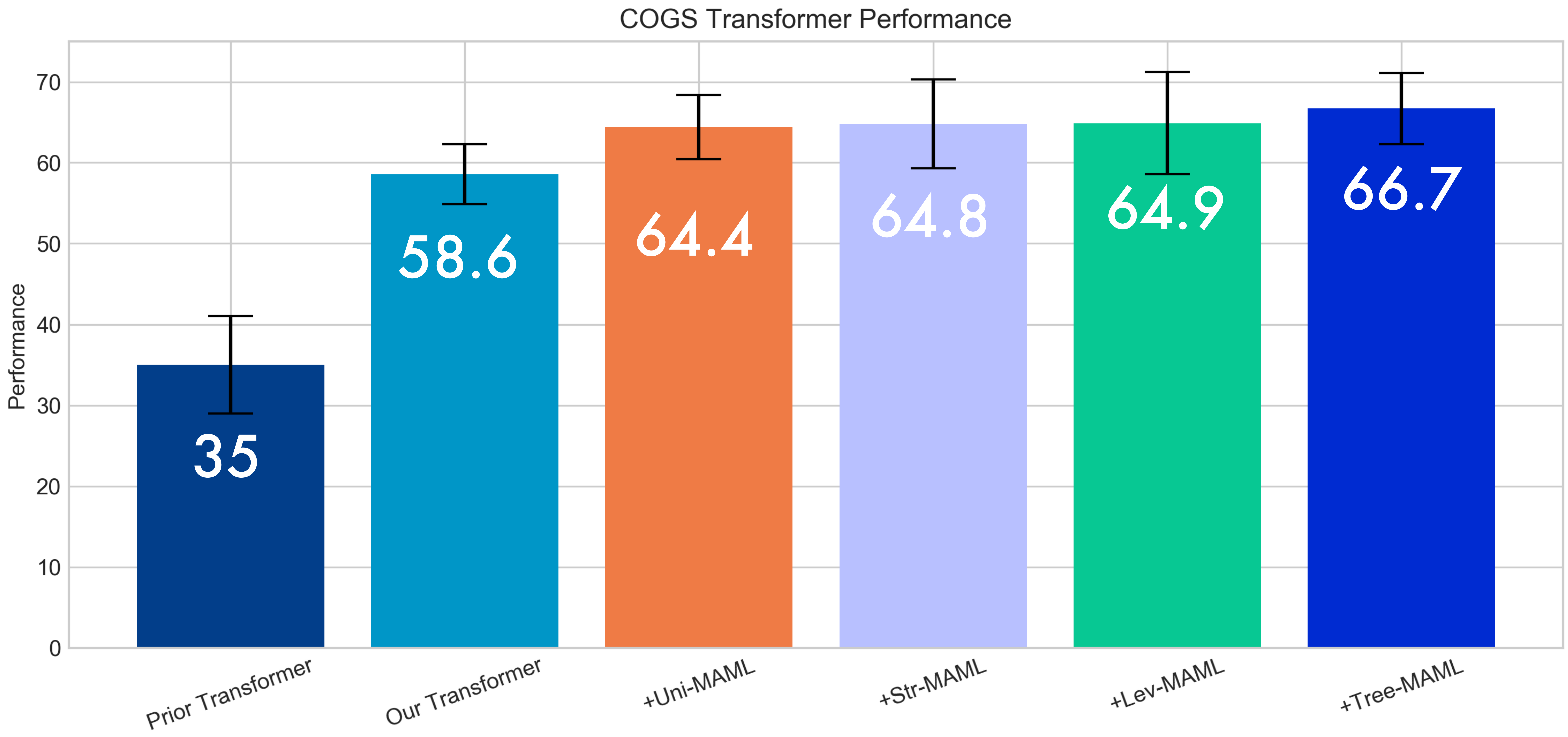
Experiments

COGS Dataset



Experiments

COGS Dataset



Conclusions

Conclusions

- Biasing the model against memorization **does improve generalization performance** on COGS and SCAN

Conclusions

- Biasing the model against memorization **does improve generalization performance** on COGS and SCAN
- Unlike more task specific methods this approach is **model** and in many cases **data-set agnostic**

Conclusions

- Biasing the model against memorization **does improve generalization performance** on COGS and SCAN

- Unlike more task specific methods this approach is **model** and in many cases **data-set agnostic**

- The design of the Meta-Test task allows for the **design of the bias applied** during training

thank you

Henry Conklin*

@henryconklin

Bailin Wang*

@bailin_28

Kenny Smith

@kennysmithed

Ivan Titov

@iatitov