Noisy Channel Language Model Prompting for Few-shot Text Classification

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




UWNLP Meta AI2

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

LM Prompting

Using a frozen language model for a downstream task

(Brown et al. 2020)

LM Prompting

Using a frozen language model for a downstream task

“Why are boolean values capitalized in Python?”

(Brown et al. 2020)

LM Prompting

Using a frozen language model for a downstream task

“Why are boolean values capitalized in Python?”

<options>

Society & Culture

Computer & Internet

Business & Finance

Entertainment & Music

Politics & Government

⋮

(Brown et al. 2020)

LM Prompting

Using a frozen language model for a downstream task

Why are boolean values capitalized in Python?

<options>

Society & Culture

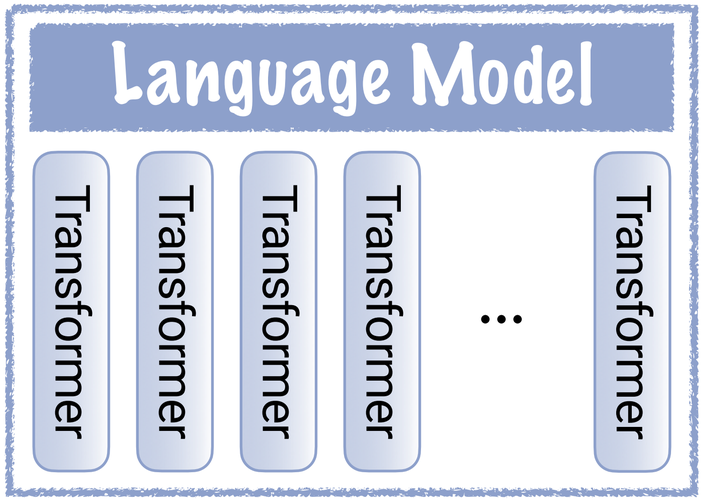
Computer & Internet

Business & Finance

Entertainment & Music

Politics & Government

⋮

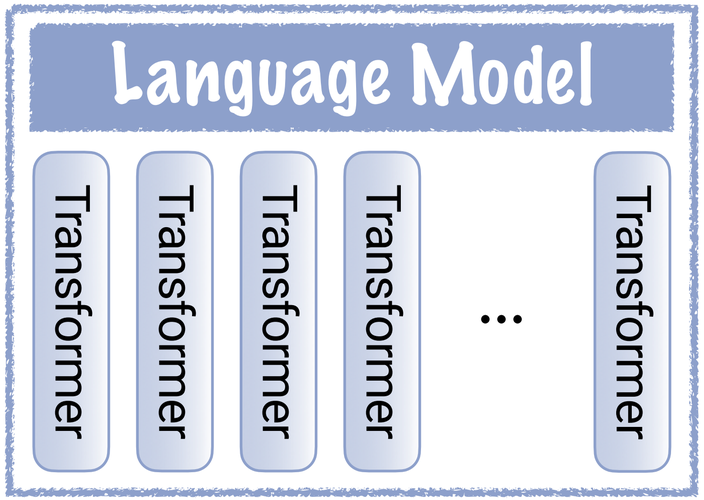


(Brown et al. 2020)

LM Prompting

Using a frozen language model for a downstream task

Why are boolean values capitalized in Python?



It is about Computer & Internet

<options>

Society & Culture
Computer & Internet
Business & Finance
Entertainment & Music
Politics & Government
⋮

(Brown et al. 2020)

LM Prompting

Using a frozen language model for a downstream task

Why are boolean values capitalized in Python?

<options>

Society & Culture

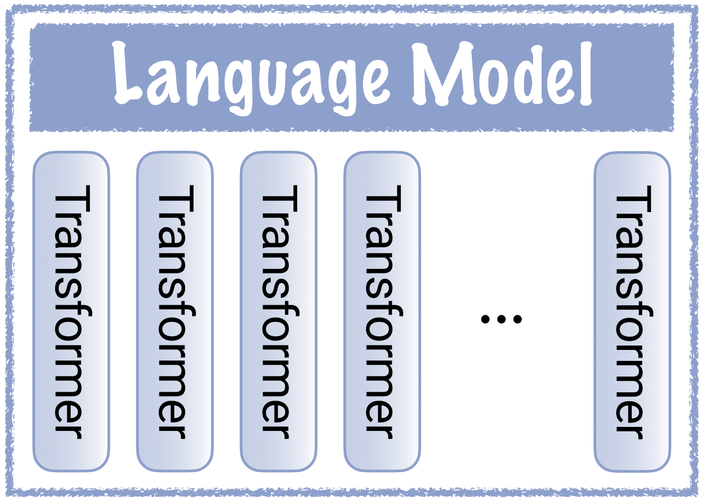
Computer & Internet

Business & Finance

Entertainment & Music

Politics & Government

⋮



It is about Computer & Internet

(Brown et al. 2020)

LM Prompting

Using a frozen language model for a downstream task

Why are boolean values capitalized in Python?

<options>

Society & Culture

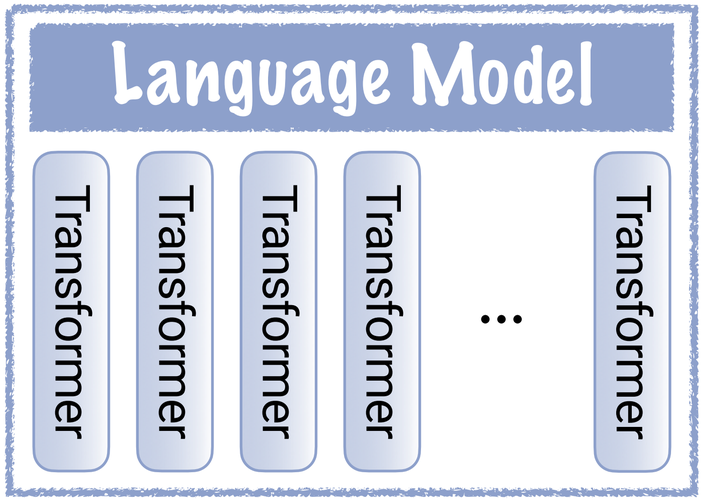
Computer & Internet

Business & Finance

Entertainment & Music

Politics & Government

⋮



It is about Computer & Internet

V

It is about Society & Culture

(Brown et al. 2020)

LM Prompting

Using a frozen language model for a downstream task

Why are boolean values capitalized in Python?

<options>

Society & Culture

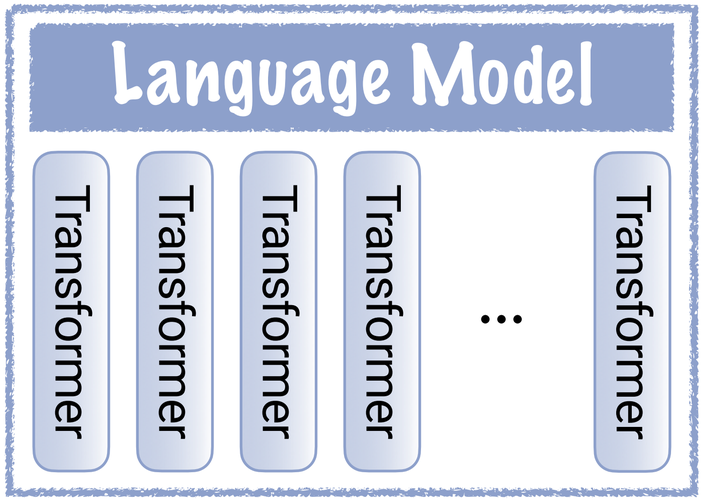
Computer & Internet ✓

Business & Finance

Entertainment & Music

Politics & Government

⋮



It is about Computer & Internet

V

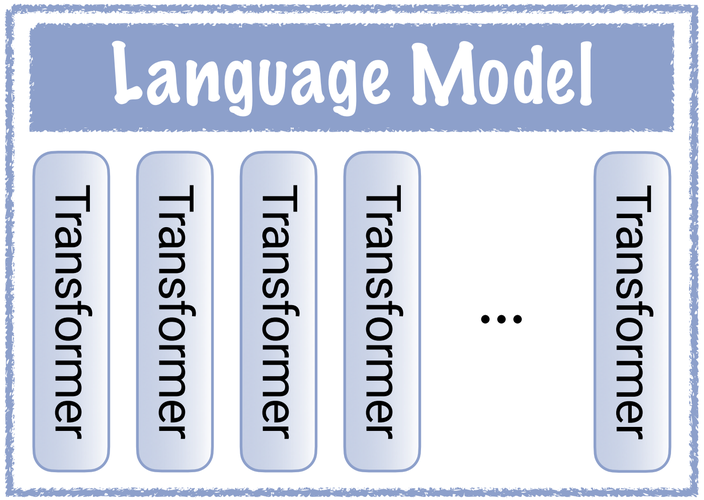
It is about Society & Culture

(Brown et al. 2020)

LM Prompting

Using a frozen language model for a downstream task

Why are boolean values capitalized in Python?



It is about Computer & Internet

V

It is about Society & Culture

<options>

Society & Culture

Computer & Internet ✓

Business & Finance

Entertainment & Music

Politics & Government

⋮

More variations

(in-context learning, prompt tuning, etc..)

(Brown et al. 2020)

LM Prompting

(Brown et al. 2020)

LM Prompting

(Brown et al. 2020)

Promising results with a 175B model

And by Zhao et al. 2021, Holtzman et al. 2021, Shin et al. 2020, Jiang et al. 2020, Gao et al. 2021, Li & Liang 2021, Lester et al. 2021, etc…

LM Prompting

(Brown et al. 2020)

Promising results with a 175B model

And by Zhao et al. 2021, Holtzman et al. 2021, Shin et al. 2020, Jiang et al. 2020, Gao et al. 2021, Li & Liang 2021, Lester et al. 2021, etc…

No or very limited number of trainable params

LM Prompting

(Brown et al. 2020)

Promising results with a 175B model

And by Zhao et al. 2021, Holtzman et al. 2021, Shin et al. 2020, Jiang et al. 2020, Gao et al. 2021, Li & Liang 2021, Lester et al. 2021, etc…

No or very limited number of trainable params

LM Prompting

GPT-3

(Brown et al. 2020)

Promising results with a 175B model

And by Zhao et al. 2021, Holtzman et al. 2021, Shin et al. 2020, Jiang et al. 2020, Gao et al. 2021, Li & Liang 2021, Lester et al. 2021, etc…

No or very limited number of trainable params

Instability — high variance

LM Prompting

(Brown et al. 2020)

Promising results with a 175B model

And by Zhao et al. 2021, Holtzman et al. 2021, Shin et al. 2020, Jiang et al. 2020, Gao et al. 2021, Li & Liang 2021, Lester et al. 2021, etc…

No or very limited number of trainable params

Instability — high variance

Low worst-case accuracy

Zhao et al. 2021, Perez et al. 2021, Lu et al. 2021

LM Prompting

(Brown et al. 2020)

Promising results with a 175B model

And by Zhao et al. 2021, Holtzman et al. 2021, Shin et al. 2020, Jiang et al. 2020, Gao et al. 2021, Li & Liang 2021, Lester et al. 2021, etc…

No or very limited number of trainable params

Instability — high variance

Low worst-case accuracy

Zhao et al. 2021, Perez et al. 2021, Lu et al. 2021

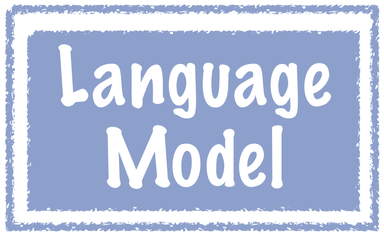
How can we do better?

Noisy Channel Approach

Input

Output

Why are boolean values capitalized in Python?



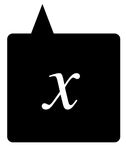
It is about Computer & Internet.

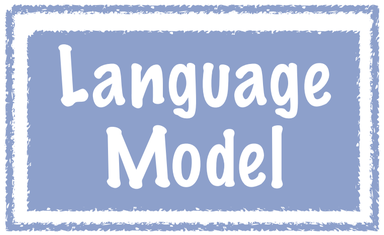
Noisy Channel Approach

Input

Output

Why are boolean values capitalized in Python?





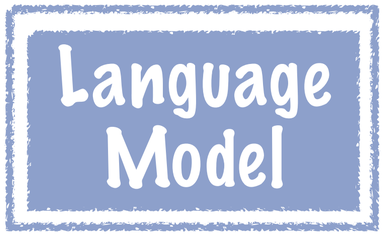
It is about Computer & Internet.

Noisy Channel Approach

Input

Output

Why are boolean values capitalized in Python?



It is about Computer & Internet.

y

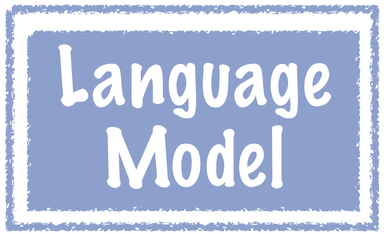
Noisy Channel Approach

Direct	$P(y \mid x)$

Input

Output

Why are boolean values capitalized in Python?



It is about Computer & Internet.

Noisy Channel Approach

Direct	$P(y \mid x)$

Input

Why are boolean values capitalized in Python?

x

Language Model

Output

It is about Computer & Internet.

y

Noisy channel model:

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)} \propto P(x \mid y)P(y)$$

Noisy Channel Approach

Input

Output

Direct $P(y|x)$

Why are boolean values capitalized in Python?

x

Language Model

It is about Computer & Internet.

y

Channel $P(x|y)P(y) \propto P(x|y)$

Noisy channel model:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \propto P(x|y)P(y)$$

Noisy Channel Approach

Input

Output

Direct $P(y|x)$

Why are boolean values capitalized in Python?

x

Language Model

It is about Computer & Internet.

y

Channel $P(x|y)P(y) \propto P(x|y)$

It is about Computer & Internet.

Language Model

Why are boolean values capitalized in Python?

Noisy channel model:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \propto P(x|y)P(y)$$

Method

x : Test input
y : Test output
$(x_1, y_1), \cdots, (x_k, y_k)$: Training data

Direct

Channel

Please see the paper for details!

Method

x : Test input
y : Test output
$(x_1, y_1), \cdots, (x_k, y_k)$: Training data

Direct

$\text{argmax}_{y \in \mathcal{Y}} P(y \mid x)$

Zero-shot

Channel

$\text{argmax}_{y \in \mathcal{Y}} P(x \mid y)$

Please see the paper for details!

Method

x : Test input
y : Test output
$(x_1, y_1), \cdots, (x_k, y_k)$: Training data

Direct		Channel
$\text{argmax}_{y \in \mathcal{Y}} P(y \mid x)$	Zero-shot	$\text{argmax}_{y \in \mathcal{Y}} P(x \mid y)$
$\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$	In-context learning	$\text{argmax}_{y \in \mathcal{Y}} P(x \mid y_1, x_1, \cdots, y_k, x_k, y)$

Please see the paper for details!

Method

x : Test input
y : Test output
$(x_1, y_1), \cdots, (x_k, y_k)$: Training data

Direct		Channel
$\text{argmax}_{y \in \mathcal{Y}} P(y \mid x)$	Zero-shot	$\text{argmax}_{y \in \mathcal{Y}} P(x \mid y)$
$\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$	In-context learning	$\text{argmax}_{y \in \mathcal{Y}} P(x \mid y_1, x_1, \cdots, y_k, x_k, y)$
$\text{argmax}_{y \in \mathcal{Y}} \Pi_{i=1}^{k} P(y \mid x_i, y_i, x)$	Ens. In-context learning	$\text{argmax}_{y \in \mathcal{Y}} \Pi_{i=1}^{k} P(x \mid y_i, x_i, y)$

Please see the paper for details!

Method

x : Test input
y : Test output
$(x_1, y_1), \cdots, (x_k, y_k)$: Training data

Direct		Channel
$\text{argmax}_{y \in \mathcal{Y}} P(y \mid x)$	Zero-shot	$\text{argmax}_{y \in \mathcal{Y}} P(x \mid y)$
$\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$	In-context learning	$\text{argmax}_{y \in \mathcal{Y}} P(x \mid y_1, x_1, \cdots, y_k, x_k, y)$
$\text{argmax}_{y \in \mathcal{Y}} \Pi_{i=1}^{k} P(y \mid x_i, y_i, x)$	Ens. In-context learning	$\text{argmax}_{y \in \mathcal{Y}} \Pi_{i=1}^{k} P(x \mid y_i, x_i, y)$
$\text{argmax}_{y \in \mathcal{Y}} P(y \mid \mathbf{V}_{\text{prompt}}, x)$	Prompt tuning (Lester et al. 2021)	$\text{argmax}_{y \in \mathcal{Y}} P(x \mid \mathbf{V}_{\text{prompt}}, y)$

($\mathbf{V}_{\text{prompt}}$ is trainable)

Please see the paper for details!

Method

x : Test input
y : Test output
$(x_1, y_1), \cdots, (x_k, y_k)$: Training data

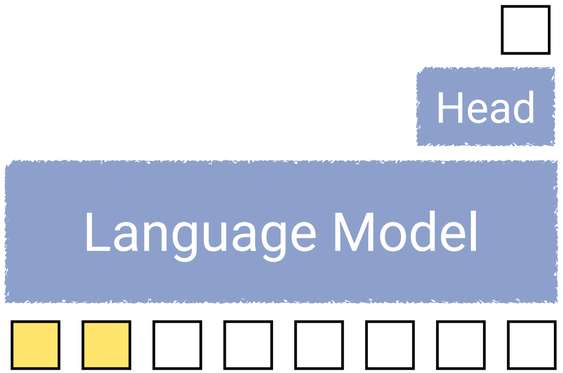
Direct		Channel
$\text{argmax}_{y \in \mathcal{Y}} P(y \mid x)$	Zero-shot	$\text{argmax}_{y \in \mathcal{Y}} P(x \mid y)$
$\text{argmax}_{y \in \mathcal{Y}} P(y \mid x_1, y_1, \cdots, x_k, y_k, x)$	In-context learning	$\text{argmax}_{y \in \mathcal{Y}} P(x \mid y_1, x_1, \cdots, y_k, x_k, y)$
$\text{argmax}_{y \in \mathcal{Y}} \Pi_{i=1}^{k} P(y \mid x_i, y_i, x)$	Ens. In-context learning	$\text{argmax}_{y \in \mathcal{Y}} \Pi_{i=1}^{k} P(x \mid y_i, x_i, y)$
$\text{argmax}_{y \in \mathcal{Y}} P(y \mid \mathbf{V}_{\text{prompt}}, x)$	Prompt tuning (Lester et al. 2021)	$\text{argmax}_{y \in \mathcal{Y}} P(x \mid \mathbf{V}_{\text{prompt}}, y)$

($\mathbf{V}_{\text{prompt}}$ is trainable)

Please see the paper for details!

Methods – Prompt tuning setup

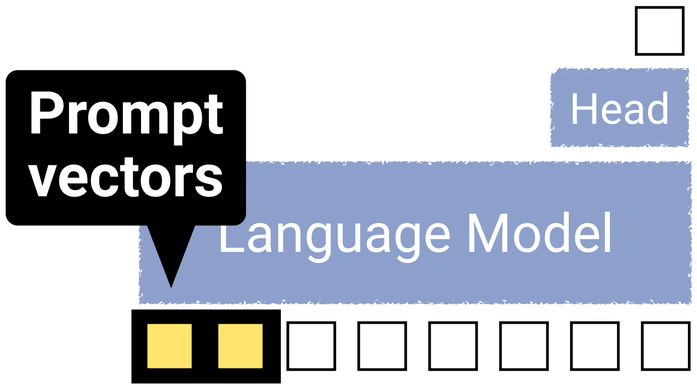
Methods – Prompt tuning setup



Prompt tuning (Lester et al. 2021)

(Yellow box = trainable)

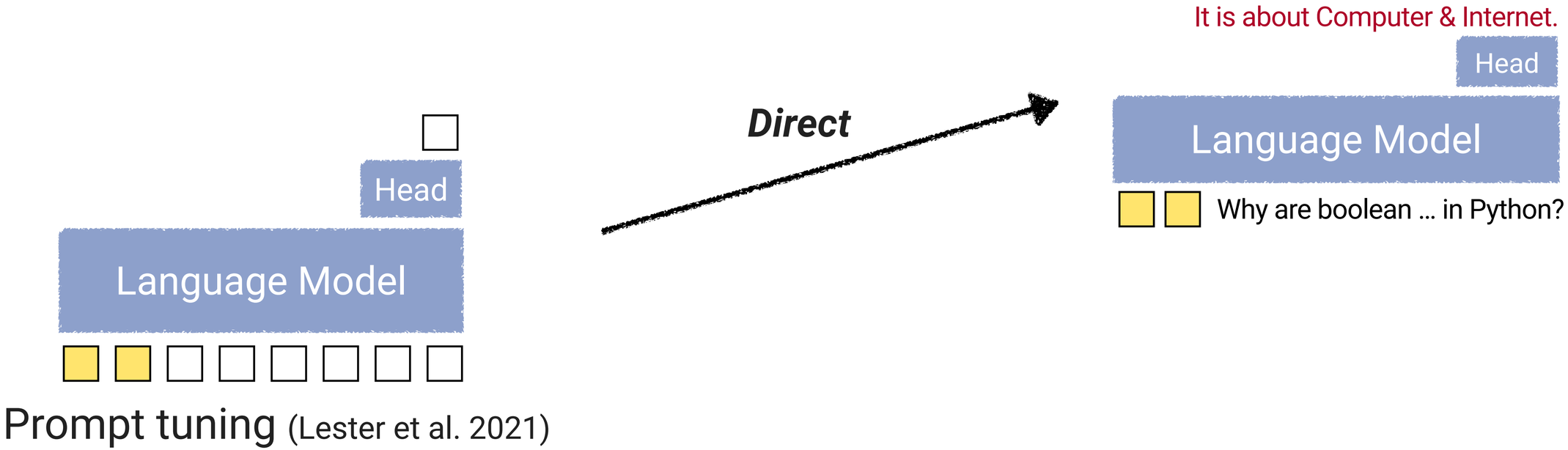
Methods – Prompt tuning setup



Prompt tuning (Lester et al. 2021)

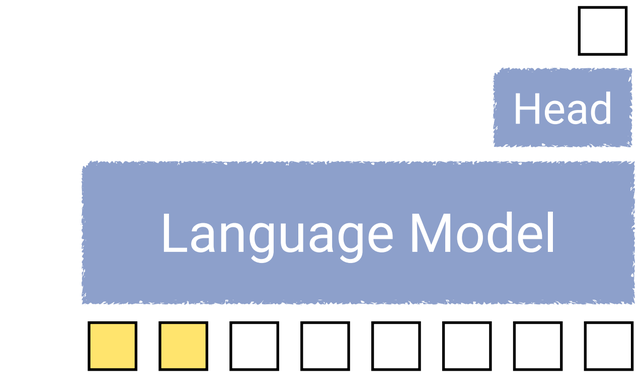
(Yellow box = trainable)

Methods – Prompt tuning setup

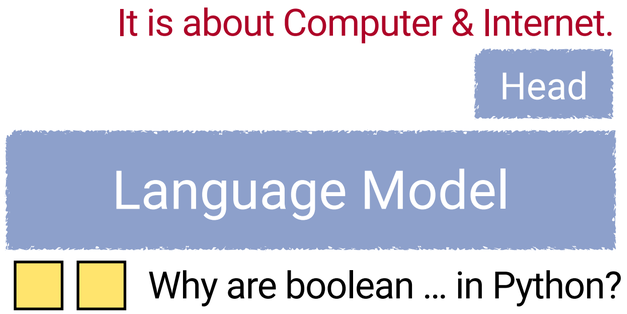


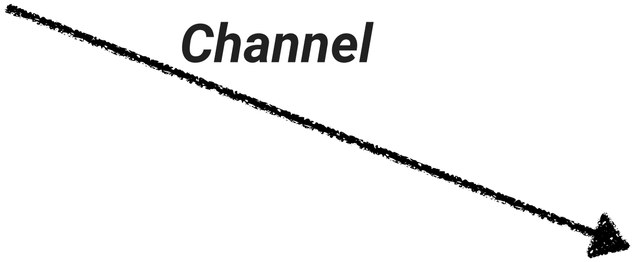
(Yellow box = trainable)

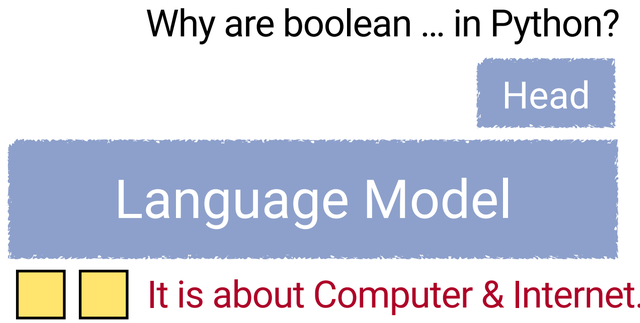
Methods – Prompt tuning setup



Prompt tuning (Lester et al. 2021)

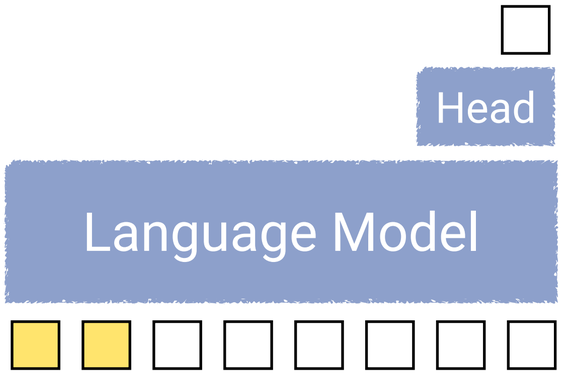






(Yellow box = trainable)

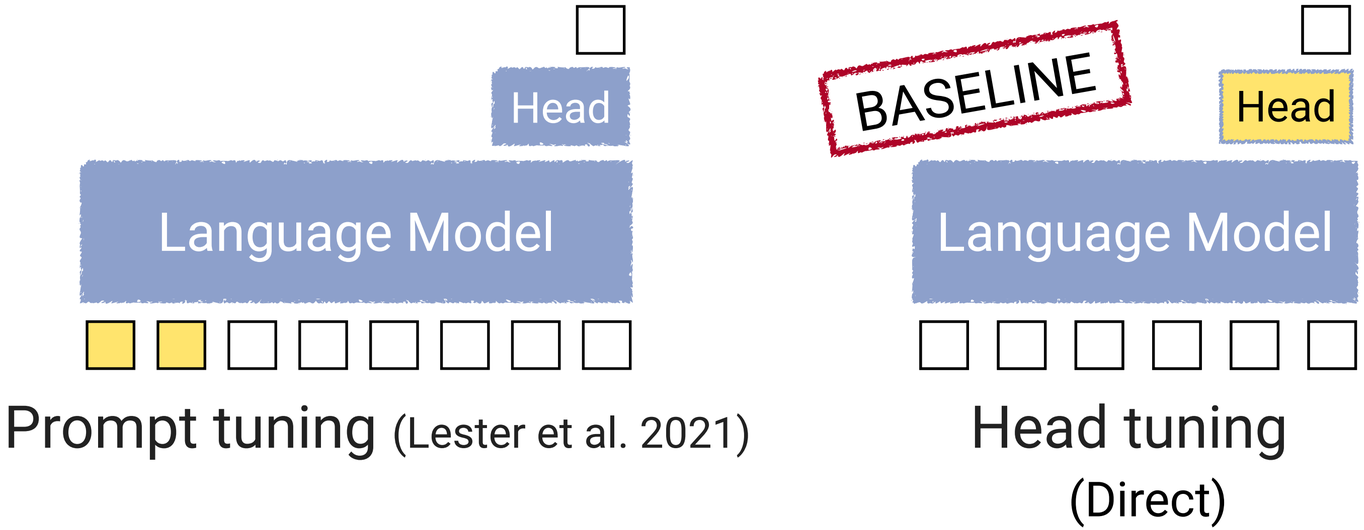
Methods – Prompt tuning setup



Prompt tuning (Lester et al. 2021)

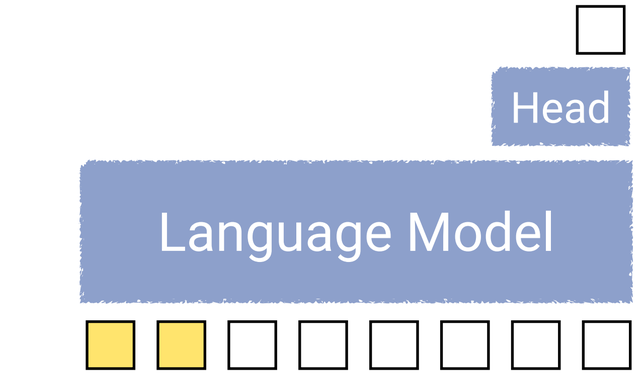
(Yellow box = trainable)

Methods – Prompt tuning setup

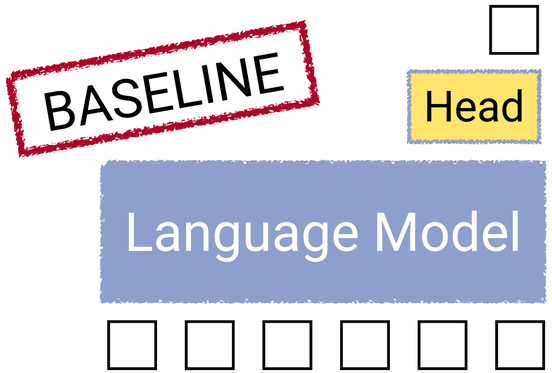


(Yellow box = trainable)

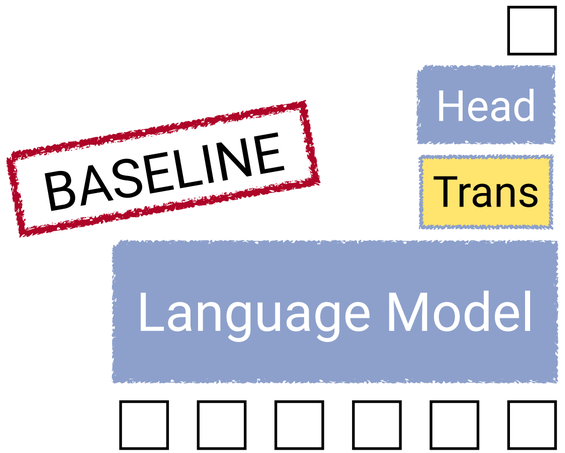
Methods – Prompt tuning setup



Prompt tuning (Lester et al. 2021)



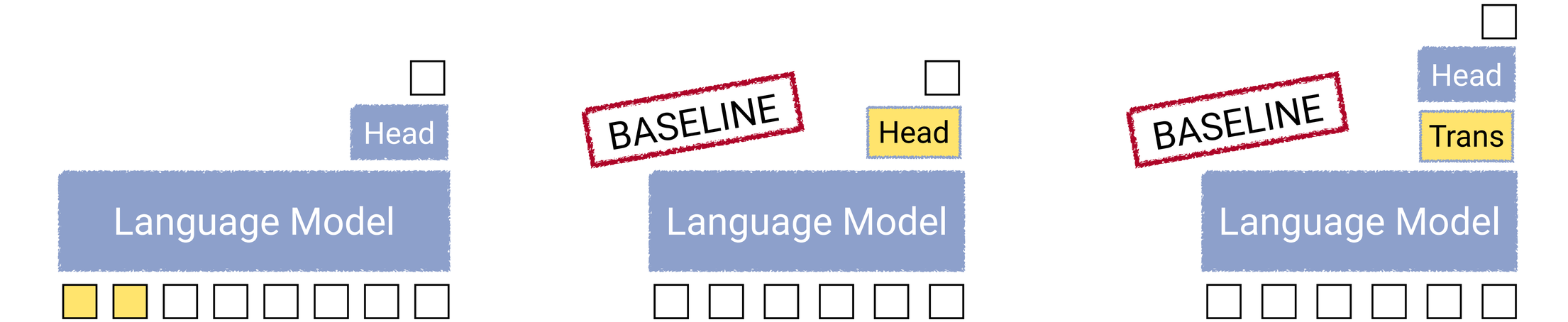
Head tuning
(Direct)



Transformation tuning
(Direct)

(Yellow box = trainable)

Methods – Prompt tuning setup



Prompt tuning (Lester et al. 2021)

Head tuning
(Direct)

Transformation tuning
(Direct)

(Yellow box = trainable)

tune ~0.002% of the total parameters

Why would it work better?

Why would it work better?

1. Few-shot learning

Noisy channel models are shown to work better in a **few-shot** setup — theoretically (Ng and Jordan, 2002) and empirically (Ding and Gimpel, 2019)

Why would it work better?

1. Few-shot learning

Noisy channel models are shown to work better in a **few-shot** setup — theoretically (Ng and Jordan, 2002) and empirically (Ding and Gimpel, 2019)

Predicting the entire input *amplifies* training signals (Lewis and Fan, 2018)

Why would it work better?

1. Few-shot learning

Noisy channel models are shown to work better in a **few-shot** setup — theoretically (Ng and Jordan, 2002) and empirically (Ding and Gimpel, 2019)

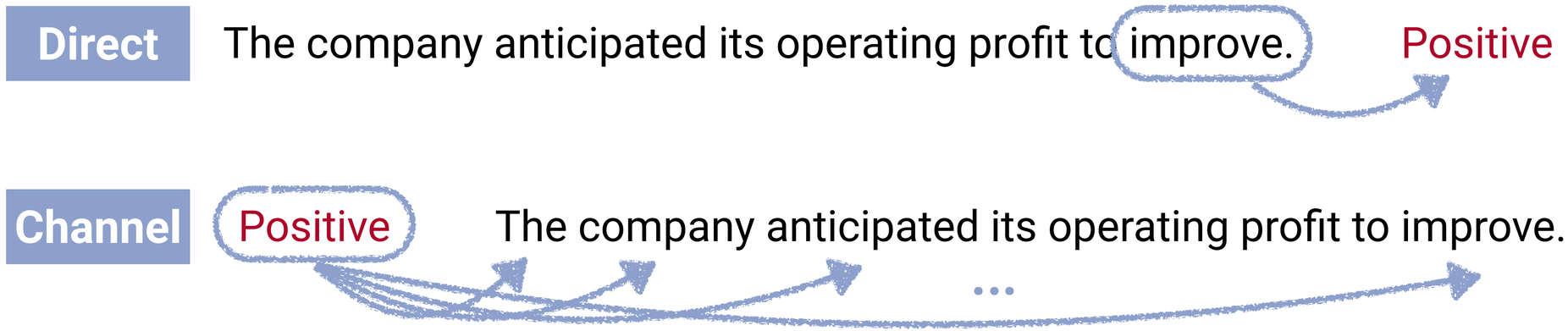
Direct The company anticipated its operating profit to improve. → Positive

Predicting the entire input *amplifies* training signals (Lewis and Fan, 2018)

Why would it work better?

1. Few-shot learning

Noisy channel models are shown to work better in a **few-shot** setup — theoretically (Ng and Jordan, 2002) and empirically (Ding and Gimpel, 2019)



Predicting the entire input *amplifies* training signals (Lewis and Fan, 2018)

Why would it work better?

2. Distribution shifts

Noisy channel models are shown to work better under **distribution shifts** (Yogtama et al. 2017, Lewis and Fan, 2018)

Why would it work better?

2. Distribution shifts

Noisy channel models are shown to work better under **distribution shifts** (Yogtama et al. 2017, Lewis and Fan, 2018)

Pre-training

(Typical) language model

Testing

(input, output) pair
on a downstream task

1) LM prompting always needs to deal with **distribution shifts**

Why would it work better?

2. Distribution shifts

Noisy channel models are shown to work better under **distribution shifts** (Yogtama et al. 2017, Lewis and Fan, 2018)

1) LM prompting always needs to deal with **distribution shifts**

2) Channel model may be better at predicting **unseen** labels

Why would it work better?

2. Distribution shifts

Noisy channel models are shown to work better under **distribution shifts** (Yogtama et al. 2017, Lewis and Fan, 2018)

$$\mathscr{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N} \qquad \{(y_i)\}_{i=1}^{N} = \{\text{positive, negative}\}$$

1) LM prompting always needs to deal with **distribution shifts**

2) Channel model may be better at predicting **unseen** labels

Why would it work better?

2. Distribution shifts

Noisy channel models are shown to work better under **distribution shifts** (Yogtama et al. 2017, Lewis and Fan, 2018)

$$\mathscr{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N} \qquad \{(y_i)\}_{i=1}^{N} = \{\text{positive, negative}\}$$

Panostaja did not disclose the purchase price. 

neutral

1) LM prompting always needs to deal with **distribution shifts**

2) Channel model may be better at predicting **unseen** labels

Why would it work better?

2. Distribution shifts

Noisy channel models are shown to work better under **distribution shifts** (Yogtama et al. 2017, Lewis and Fan, 2018)

$$\mathscr{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N} \qquad \{(y_i)\}_{i=1}^{N} = \{\text{positive, negative}\}$$

Panostaja did not disclose the purchase price. 

neutral ←*likely very low*

1) LM prompting always needs to deal with **distribution shifts**

2) Channel model may be better at predicting **unseen** labels

Why would it work better?

2. Distribution shifts

Noisy channel models are shown to work better under **distribution shifts** (Yogtama et al. 2017, Lewis and Fan, 2018)

$$\mathscr{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N} \qquad \{(y_i)\}_{i=1}^{N} = \{\text{positive, negative}\}$$

Panostaja did not disclose the purchase price. LM neutral ←*likely very low*

neutral LM Panostaja did not disclose the purchase price. ←*?*

1) LM prompting always needs to deal with **distribution shifts**

2) Channel model may be better at predicting **unseen** labels

Why would it work better?

2. Distribution shifts

Noisy channel models are shown to work better under **distribution shifts** (Yogtama et al. 2017, Lewis and Fan, 2018)

$$\mathscr{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N} \qquad \{(y_i)\}_{i=1}^{N} = \{\text{positive, negative}\}$$

Panostaja did not disclose the purchase price. LM neutral ←*likely very low*

neutral LM Panostaja did not disclose the purchase price. ←*?*

1) LM prompting always needs to deal with **distribution shifts**

2) Channel model may be better at predicting **unseen** labels

(experiments showing this later!)

Experimental Setup

Dataset	Task	# labels
SST-2	Sentiment analysis (movie)	2
SST-5	Sentiment analysis (movie)	5
MR	Sentiment analysis (movie)	2
CR	Sentiment analysis (electronics)	2
Amazon	Sentiment analysis (Amazon)	5
Yelp	Sentiment analysis (Yelp)	5
TREC	Question classification (answer type)	6
AGNews	News classification (topic)	4
Yahoo	Question classification (topic)	10
DBPedia	Ontology classification	14
Subj	Subjectivity classification	2

Experimental Setup

Dataset	Task	# labels
SST-2	Sentiment analysis (movie)	2
SST-5	Sentiment analysis (movie)	5
MR	Sentiment analysis (movie)	2
CR	Sentiment analysis (electronics)	2
Amazon	Sentiment analysis (Amazon)	5
Yelp	Sentiment analysis (Yelp)	5
TREC	Question classification (answer type)	6
AGNews	News classification (topic)	4
Yahoo	Question classification (topic)	10
DBPedia	Ontology classification	14
Subj	Subjectivity classification	2

GPT-2
LARGE

(# params = 770M)

Experimental Setup

Dataset	Task	# labels
SST-2	Sentiment analysis (movie)	2
SST-5	Sentiment analysis (movie)	5
MR	Sentiment analysis (movie)	2
CR	Sentiment analysis (electronics)	2
Amazon	Sentiment analysis (Amazon)	5
Yelp	Sentiment analysis (Yelp)	5
TREC	Question classification (answer type)	6
AGNews	News classification (topic)	4
Yahoo	Question classification (topic)	10
DBPedia	Ontology classification	14
Subj	Subjectivity classification	2

GPT-2 LARGE (# params = 770M)

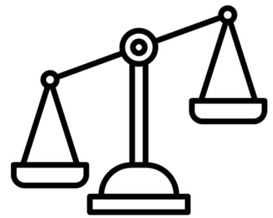
Training data: K=16

Experimental Setup

Dataset	Task	# labels
SST-2	Sentiment analysis (movie)	2
SST-5	Sentiment analysis (movie)	5
MR	Sentiment analysis (movie)	2
CR	Sentiment analysis (electronics)	2
Amazon	Sentiment analysis (Amazon)	5
Yelp	Sentiment analysis (Yelp)	5
TREC	Question classification (answer type)	6
AGNews	News classification (topic)	4
Yahoo	Question classification (topic)	10
DBPedia	Ontology classification	14
Subj	Subjectivity classification	2

GPT-2 LARGE (# params = 770M)

Training data: K=16



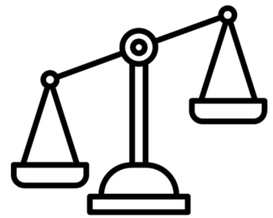
No Guarantee

Experimental Setup

Dataset	Task	# labels
SST-2	Sentiment analysis (movie)	2
SST-5	Sentiment analysis (movie)	5
MR	Sentiment analysis (movie)	2
CR	Sentiment analysis (electronics)	2
Amazon	Sentiment analysis (Amazon)	5
Yelp	Sentiment analysis (Yelp)	5
TREC	Question classification (answer type)	6
AGNews	News classification (topic)	4
Yahoo	Question classification (topic)	10
DBPedia	Ontology classification	14
Subj	Subjectivity classification	2

GPT-2 LARGE (# params = 770M)

Training data: K=16



No Guarantee

80 different runs

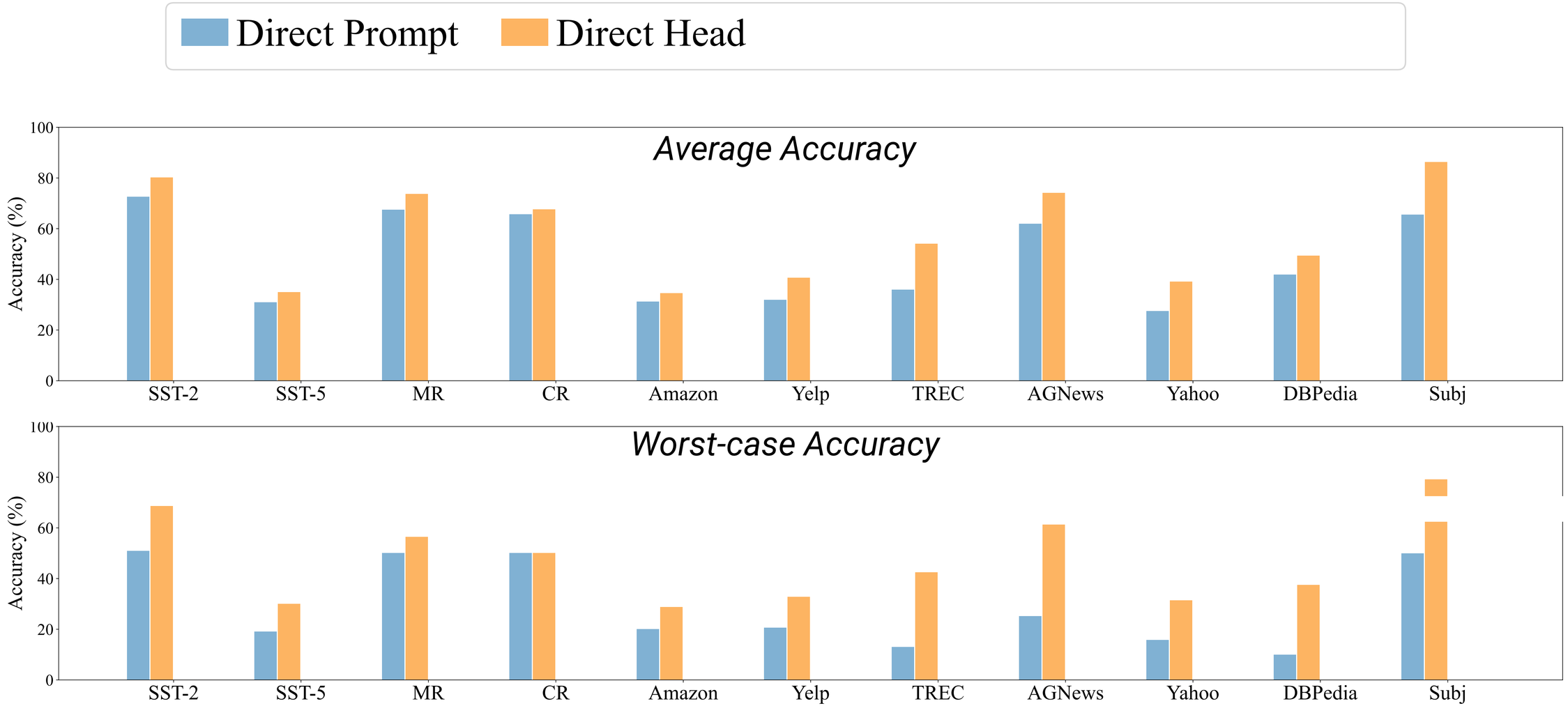
(4 x 

, 5 x 
, 4 x 
)

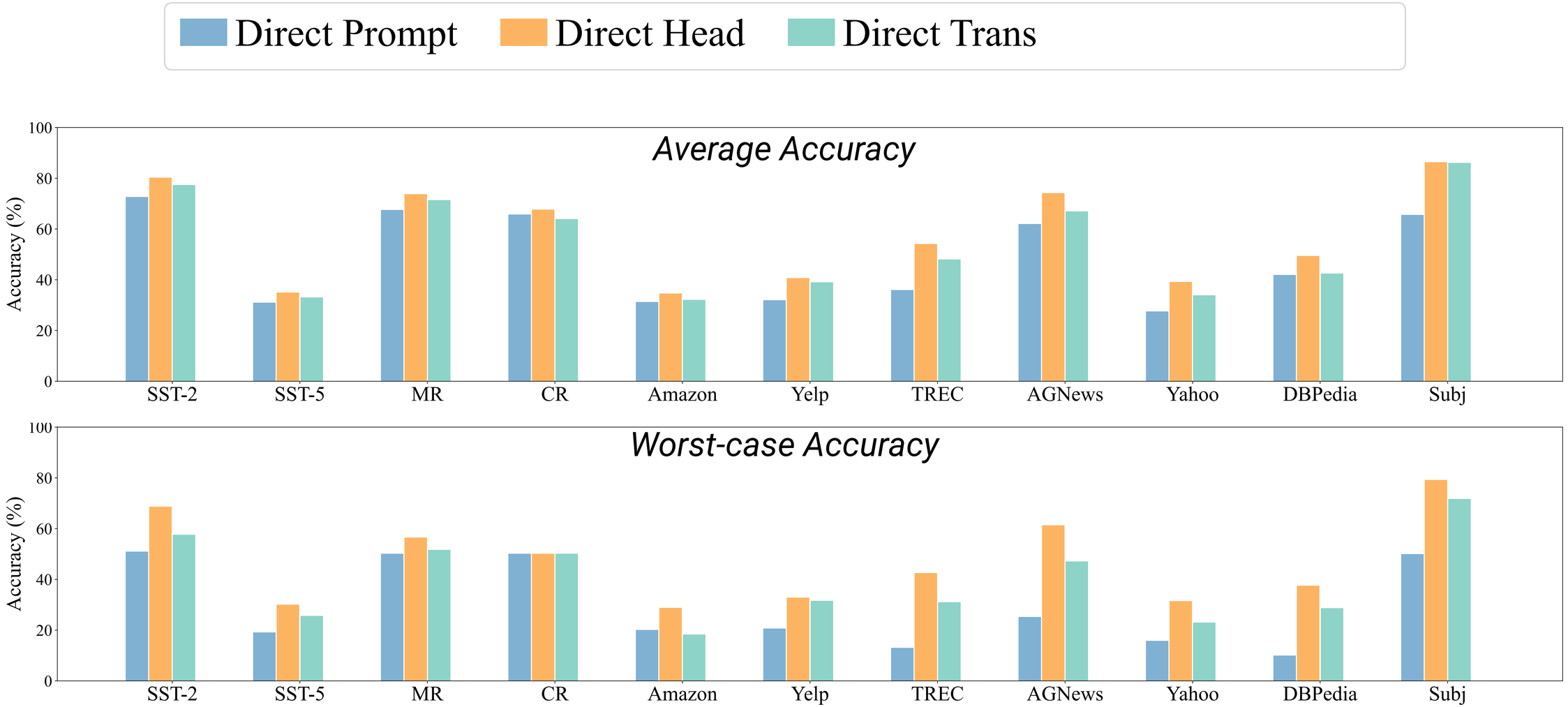
label words *data* *train*

Results – Prompt tuning setup

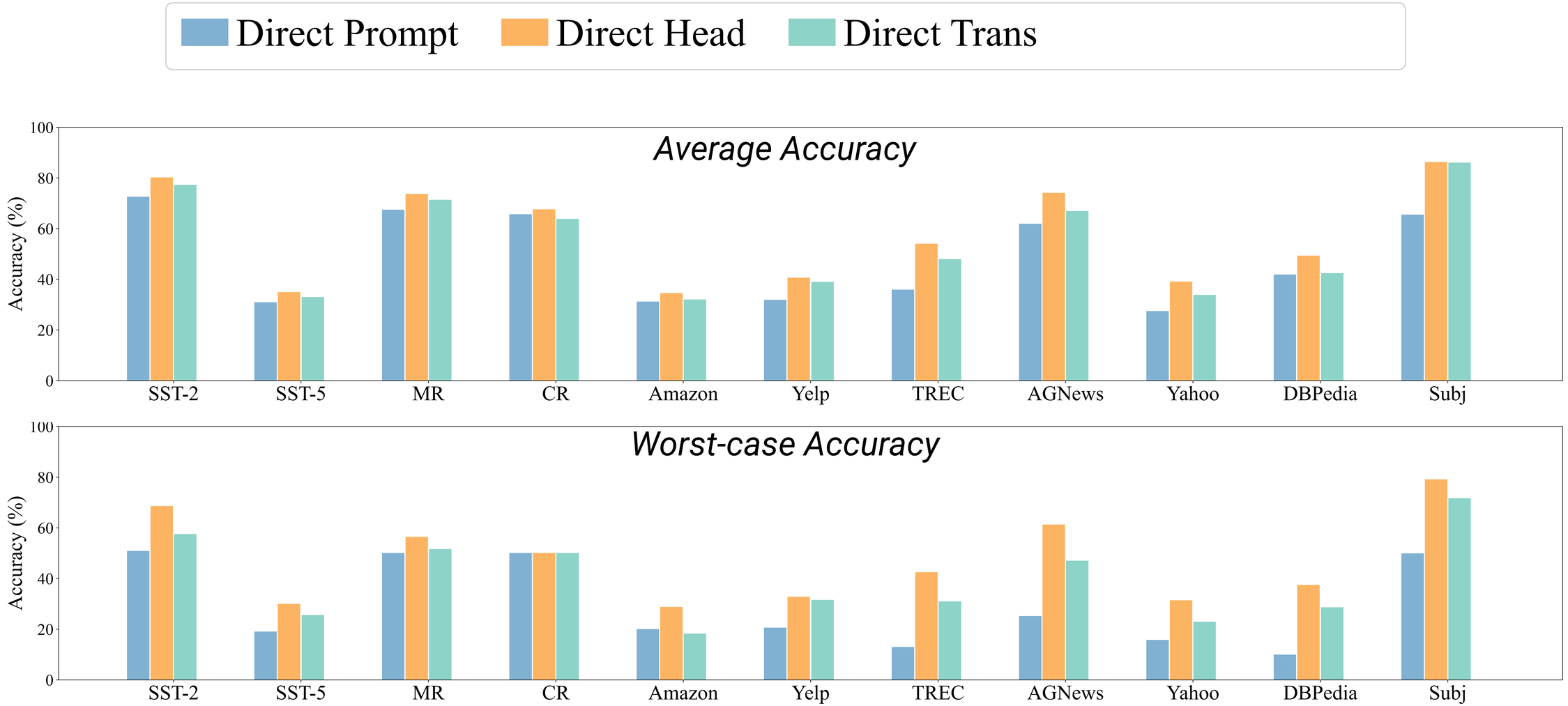
Results – Prompt tuning setup



Results – Prompt tuning setup

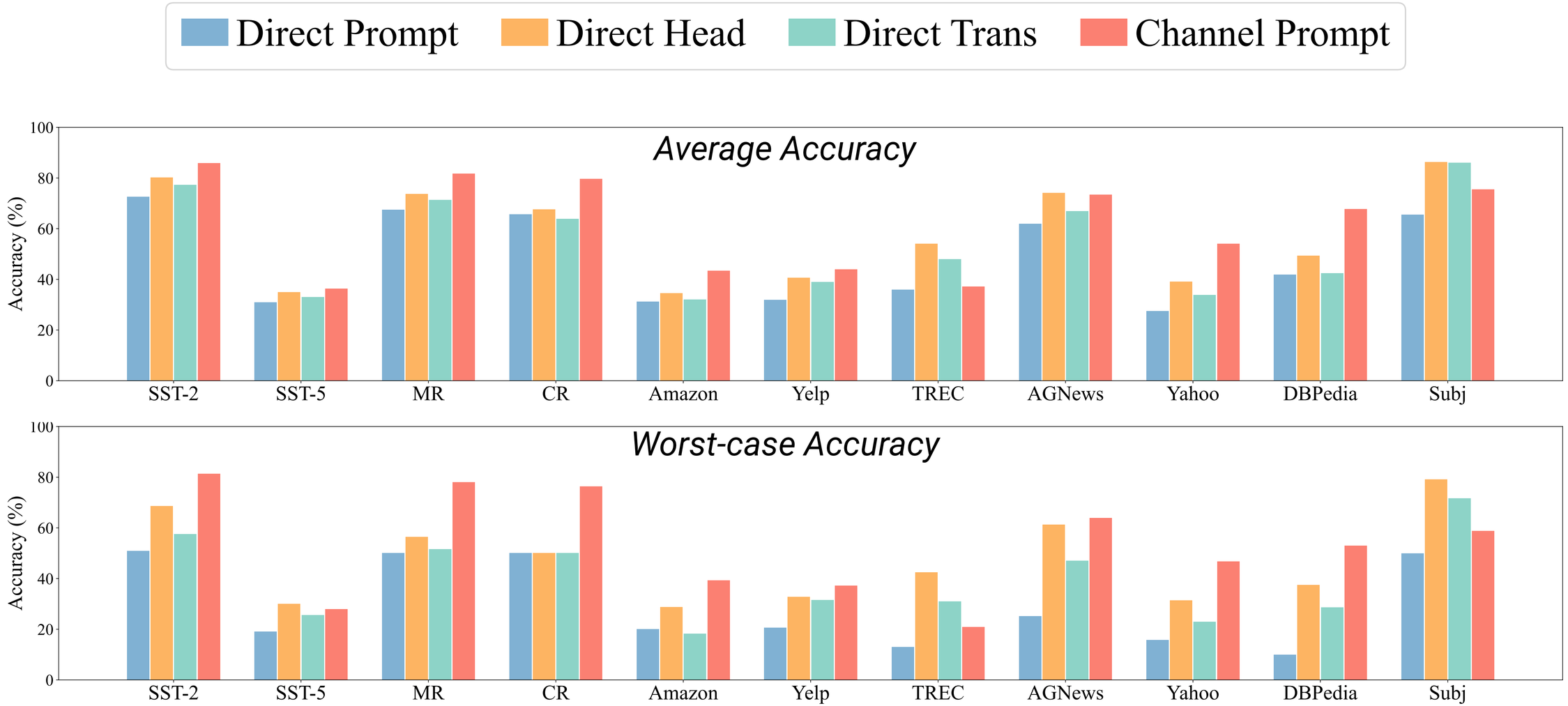


Results – Prompt tuning setup



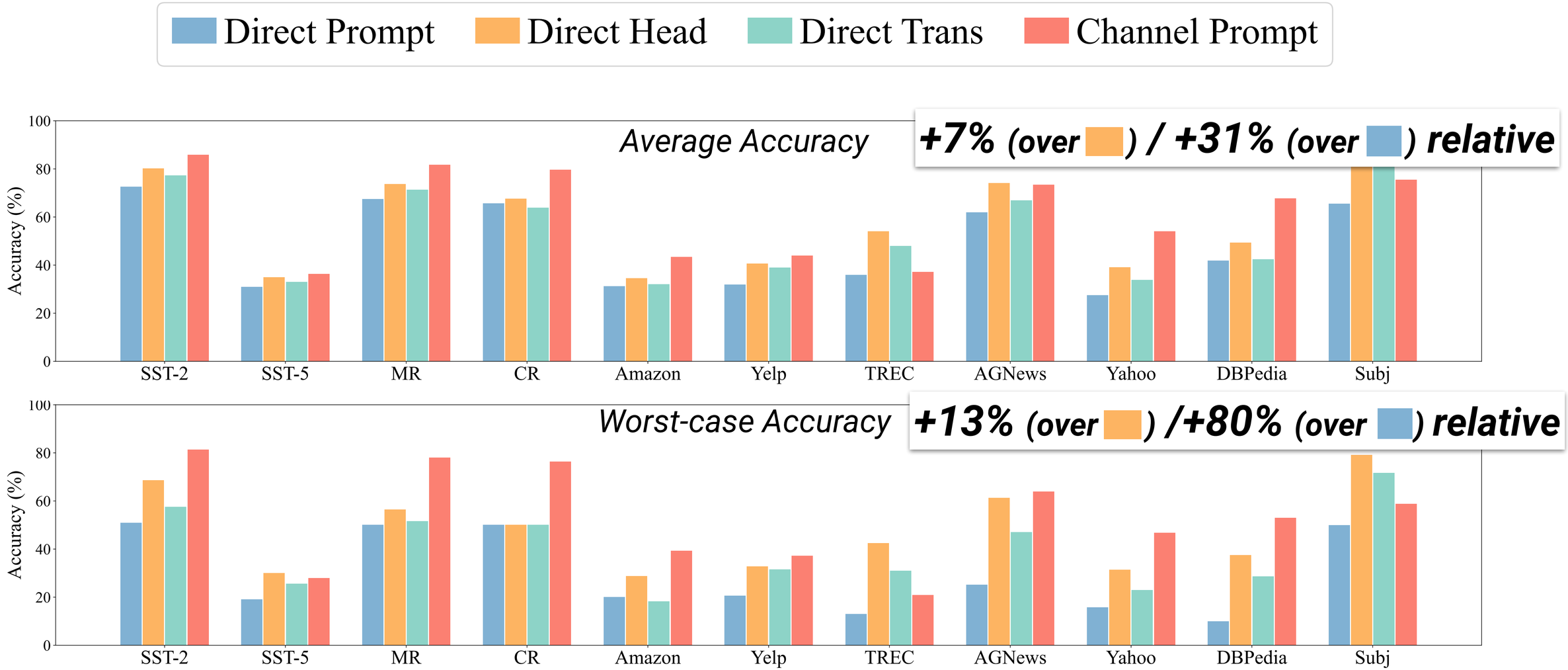
Direct Head Tuning is a powerful baseline

Results – Prompt tuning setup



Channel Prompt Tuning outperforms Direct models

Results – Prompt tuning setup



Channel Prompt Tuning outperforms Direct models
(more significantly in worst-case accuracy)

Ablations

Impact of imbalance in training data

Ablations

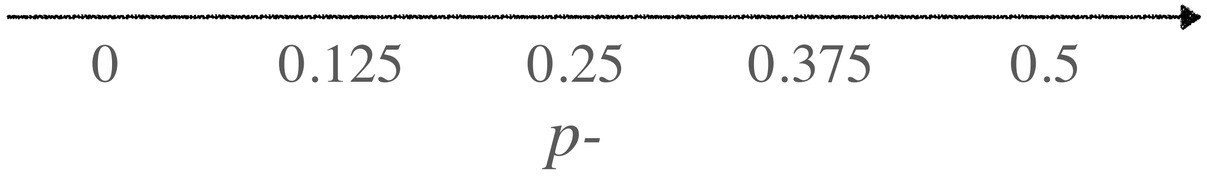
Impact of imbalance in training data

Binary classification datasets

Ablations

Impact of imbalance in training data

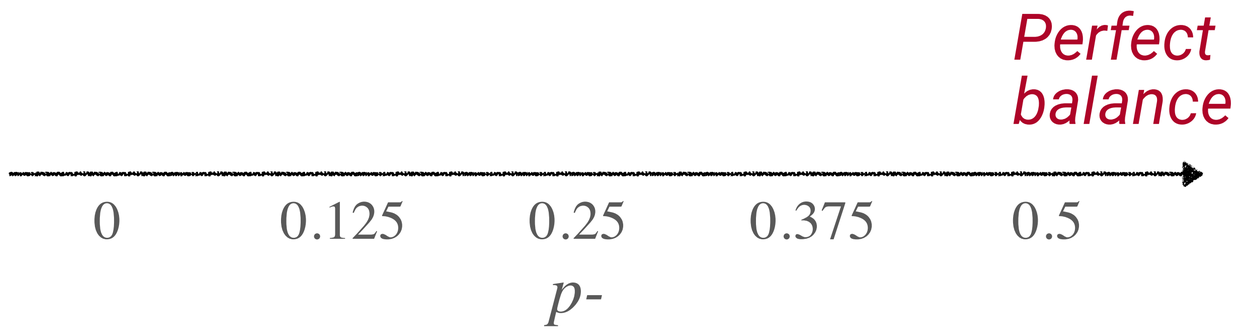
Binary classification datasets



Ablations

Impact of imbalance in training data

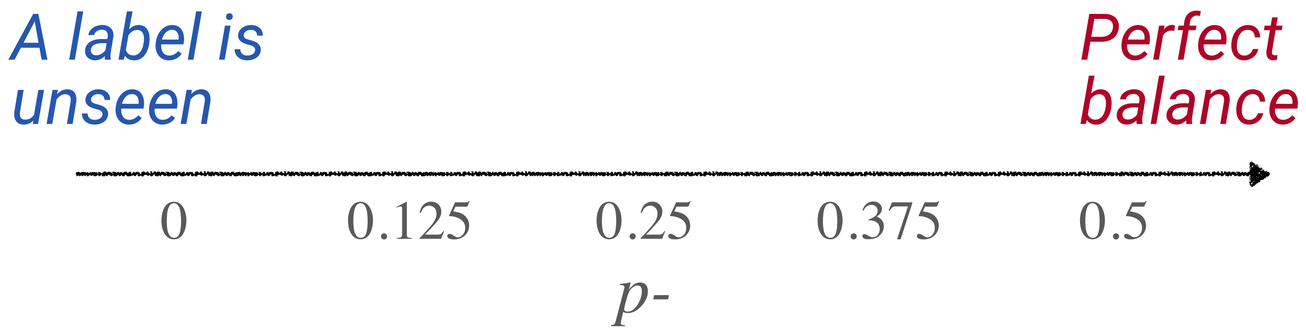
Binary classification datasets



Ablations

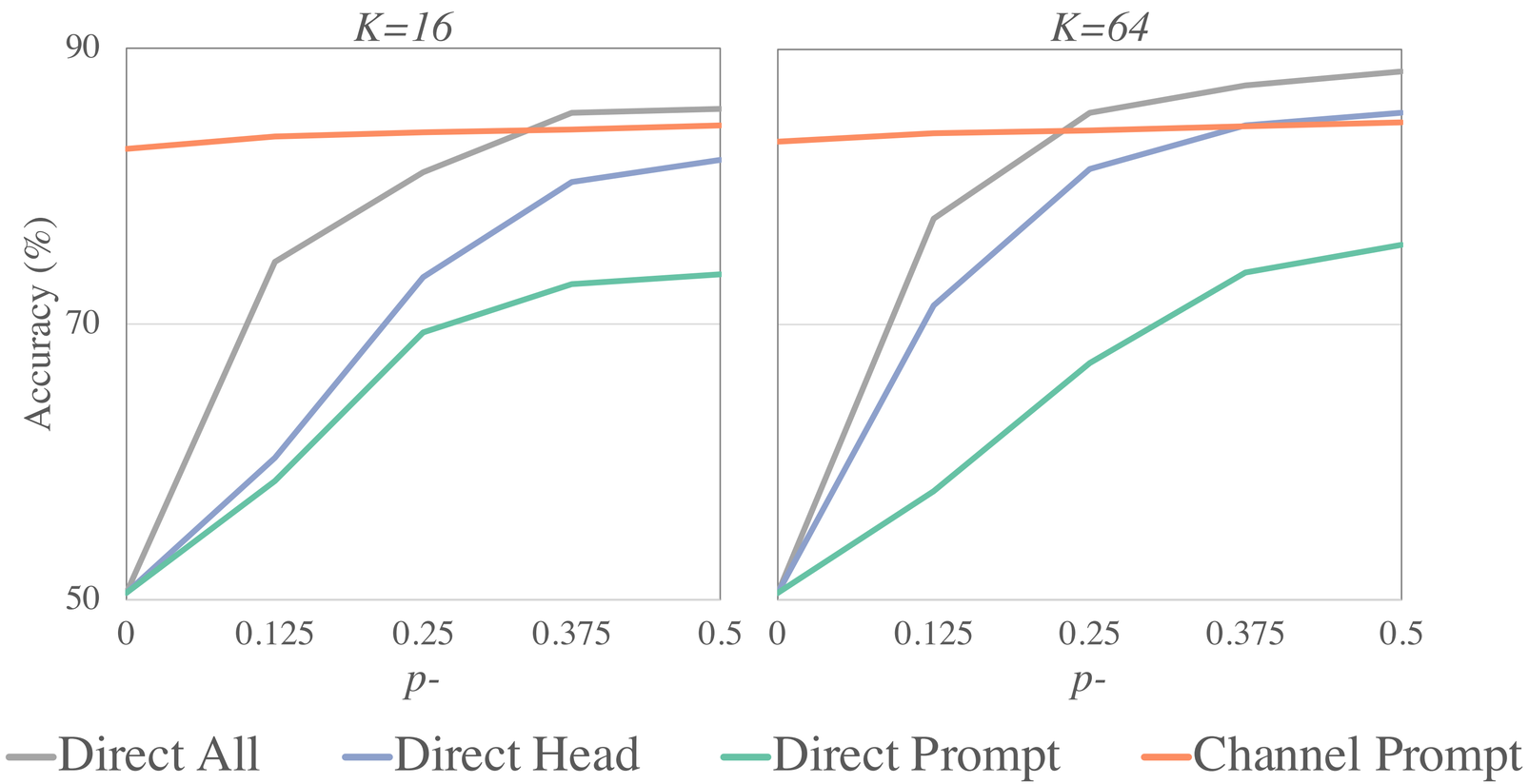
Impact of imbalance in training data

Binary classification datasets



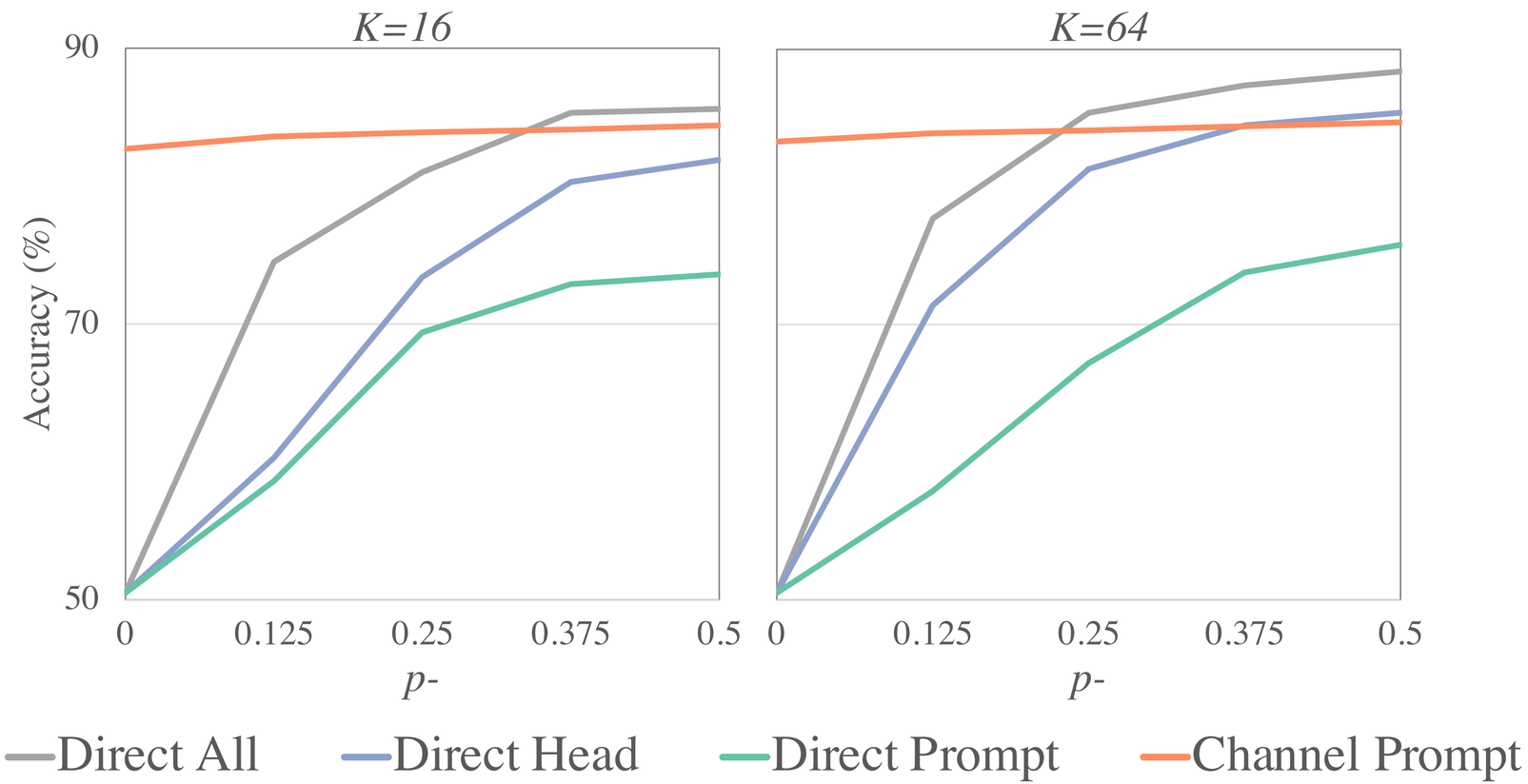
Ablations

Impact of imbalance in training data



Ablations

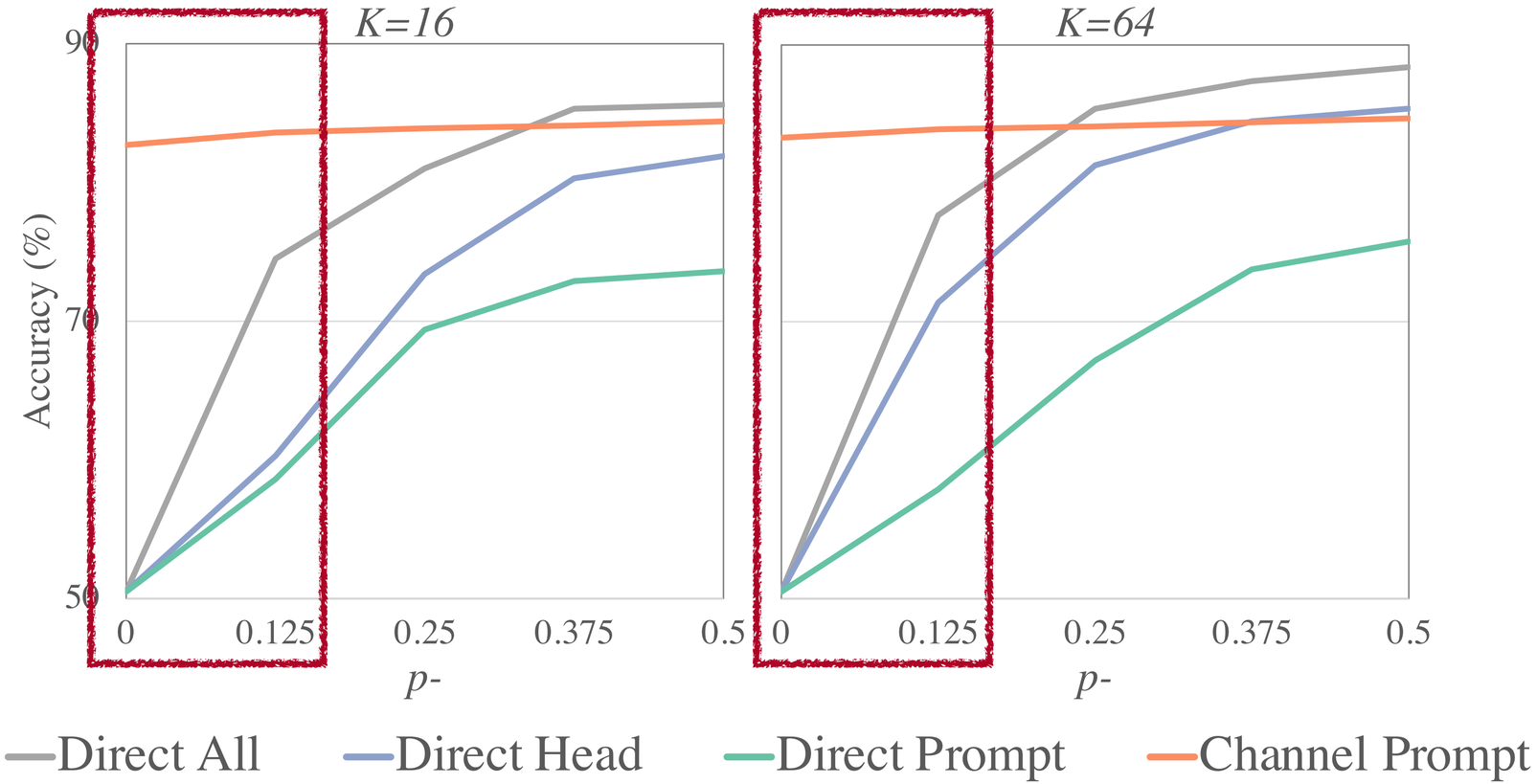
Impact of imbalance in training data



Channel models are much less sensitive to imbalance in training data

Ablations

Impact of imbalance in training data

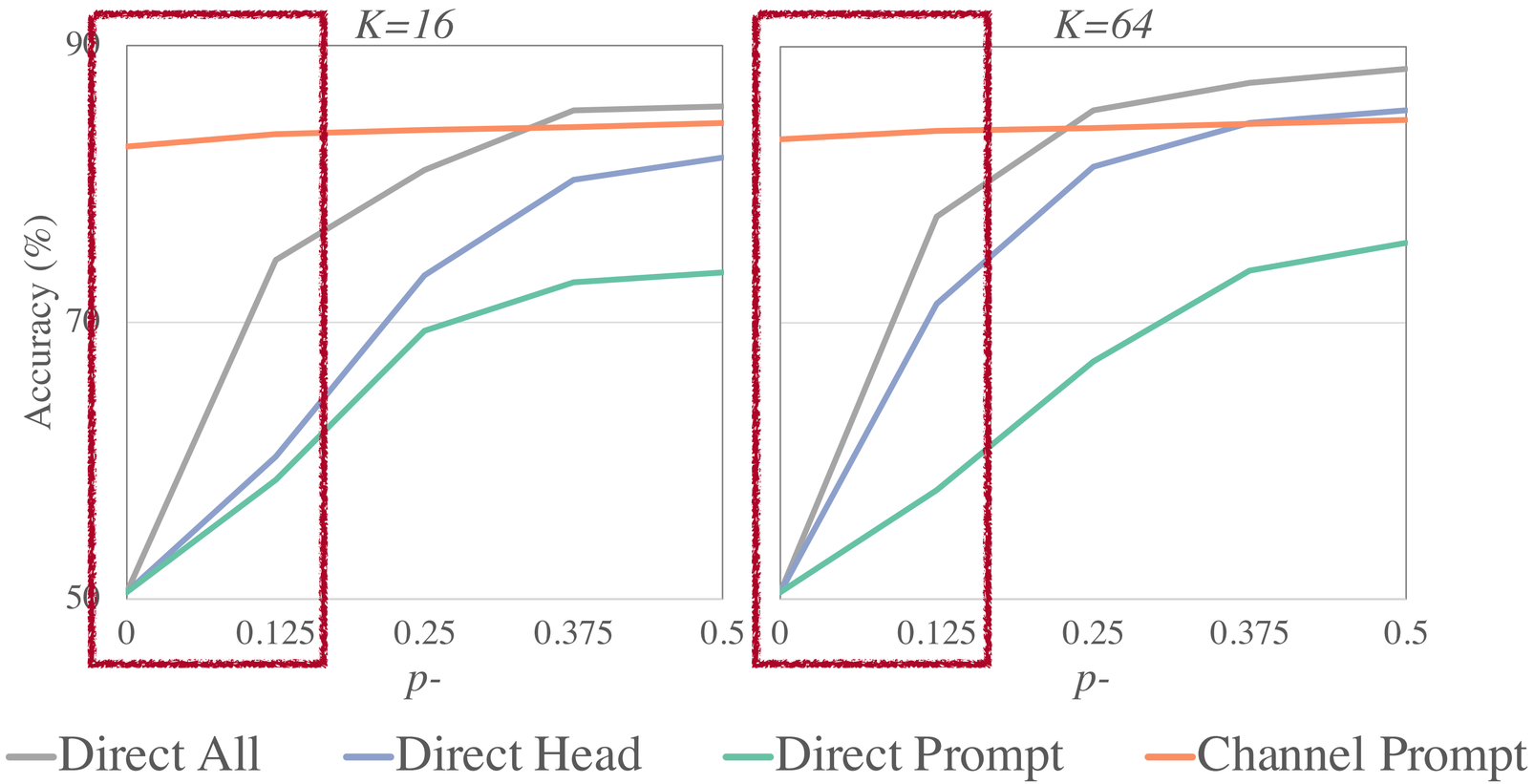


Channel models are much less sensitive to imbalance in training data

Channel models do well even with unseen labels

Ablations

Impact of imbalance in training data



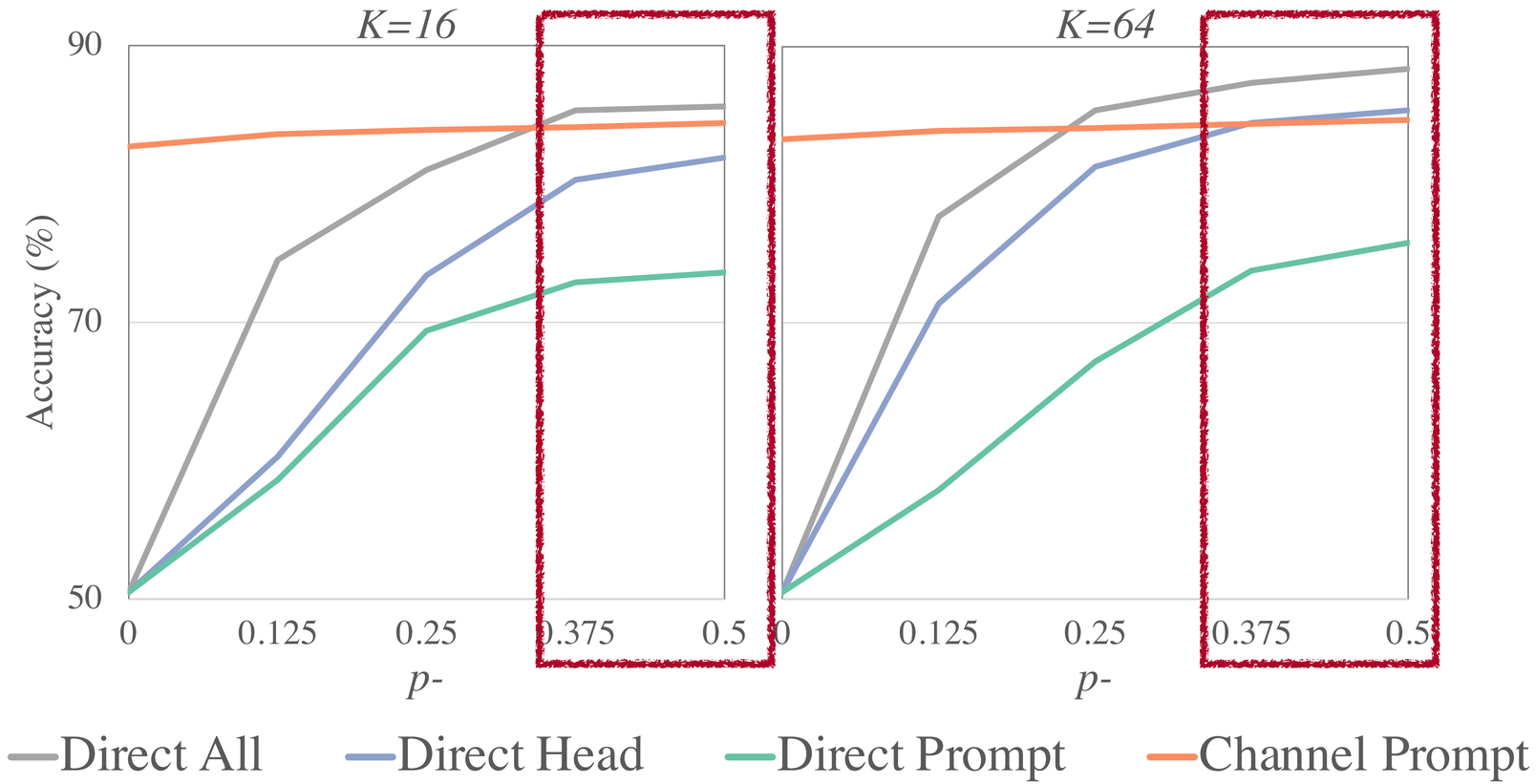
Channel models are much less sensitive to imbalance in training data

Channel models do well even with unseen labels

(Likely because labels are conditioning variables)

Ablations

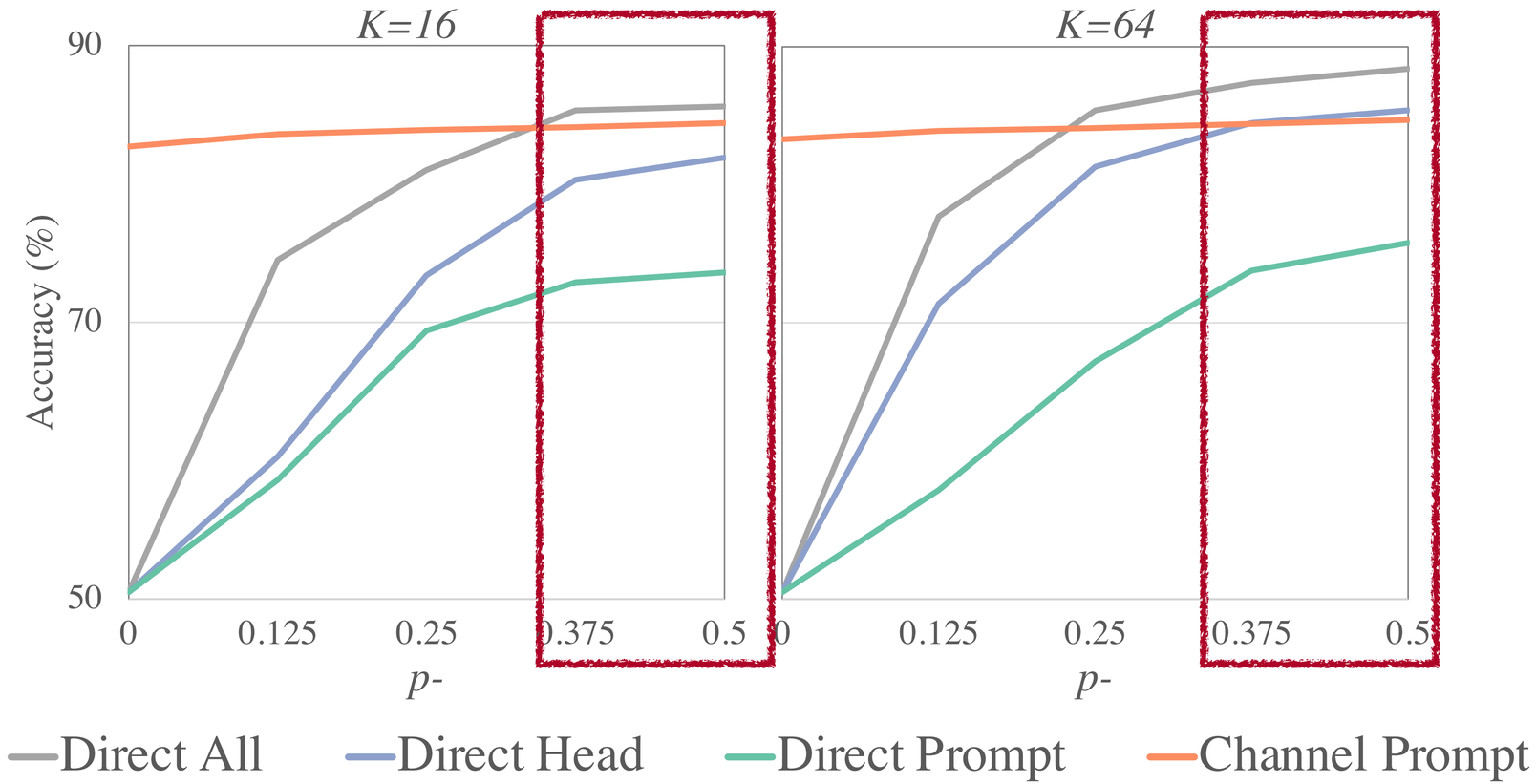
Impact of imbalance in training data



Direct models can be better with perfect balanced data

Ablations

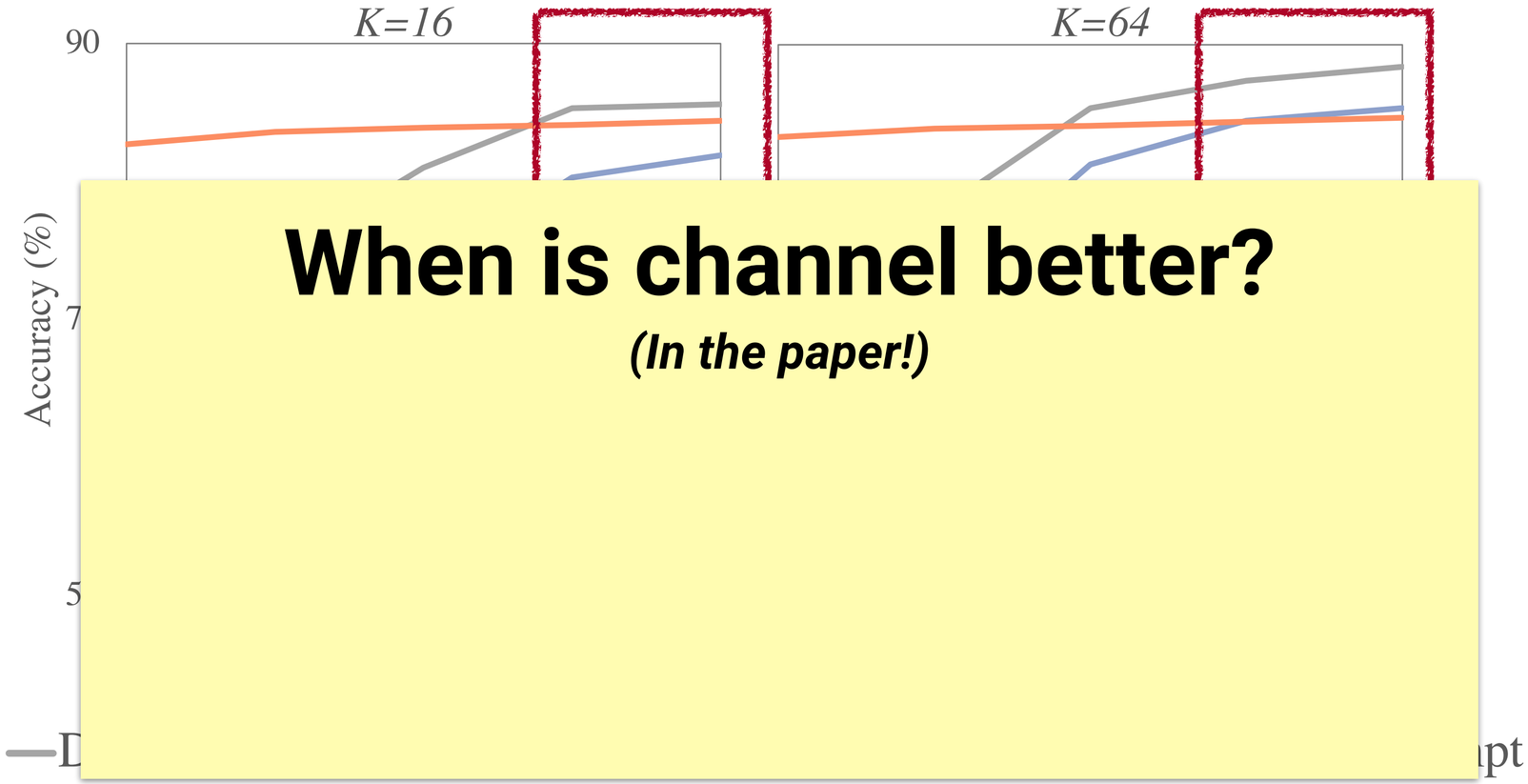
Impact of imbalance in training data



Direct models can be better with perfect balanced data

Ablations

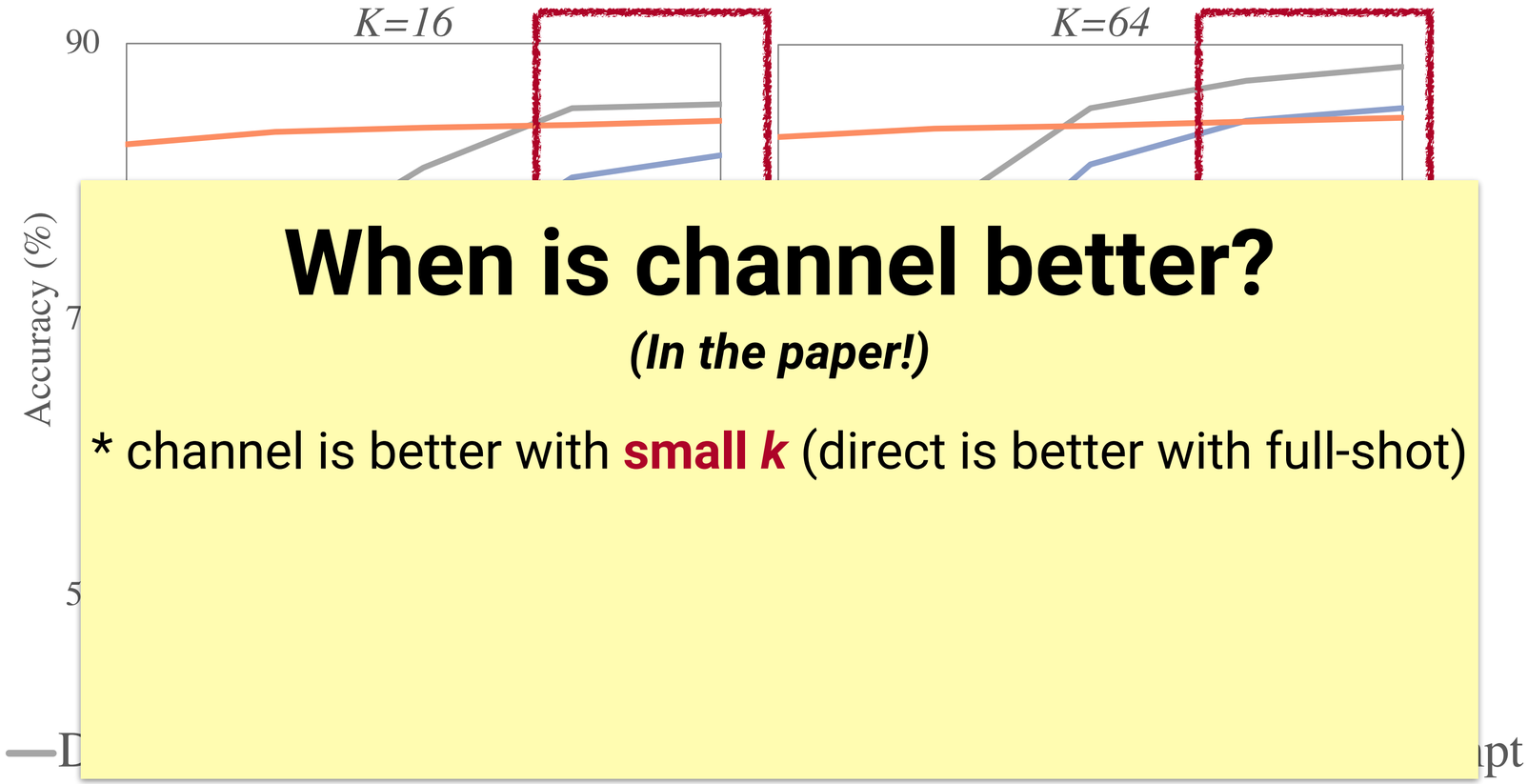
Impact of imbalance in training data



Direct models can be better with perfect balanced data

Ablations

Impact of imbalance in training data



When is channel better?

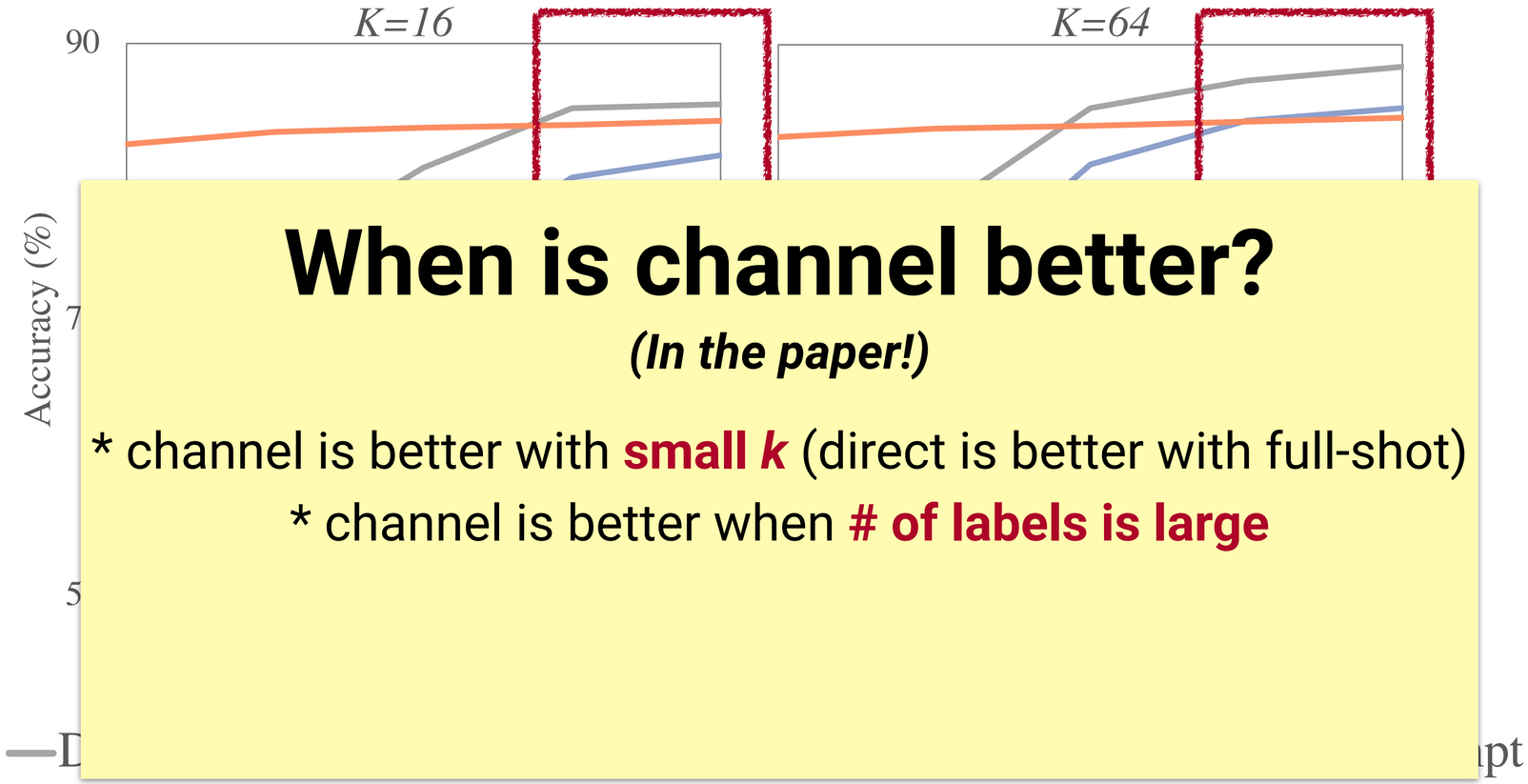
(In the paper!)

* channel is better with **small *k*** (direct is better with full-shot)

Direct models can be better with perfect balanced data

Ablations

Impact of imbalance in training data



When is channel better?

(In the paper!)

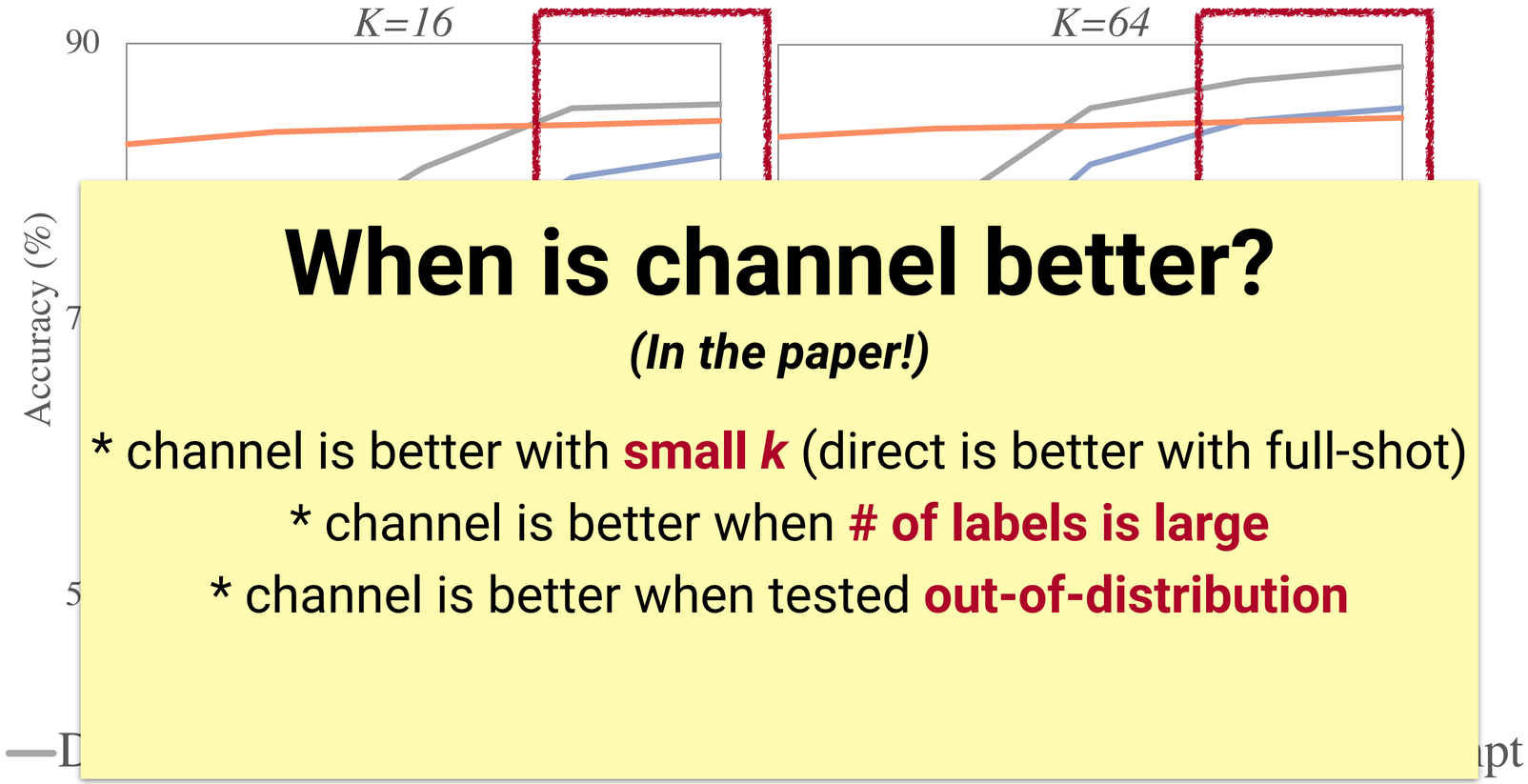
* channel is better with **small *k*** (direct is better with full-shot)

* channel is better when **# of labels is large**

Direct models can be better with perfect balanced data

Ablations

Impact of imbalance in training data



When is channel better?

(In the paper!)

* channel is better with **small *k*** (direct is better with full-shot)

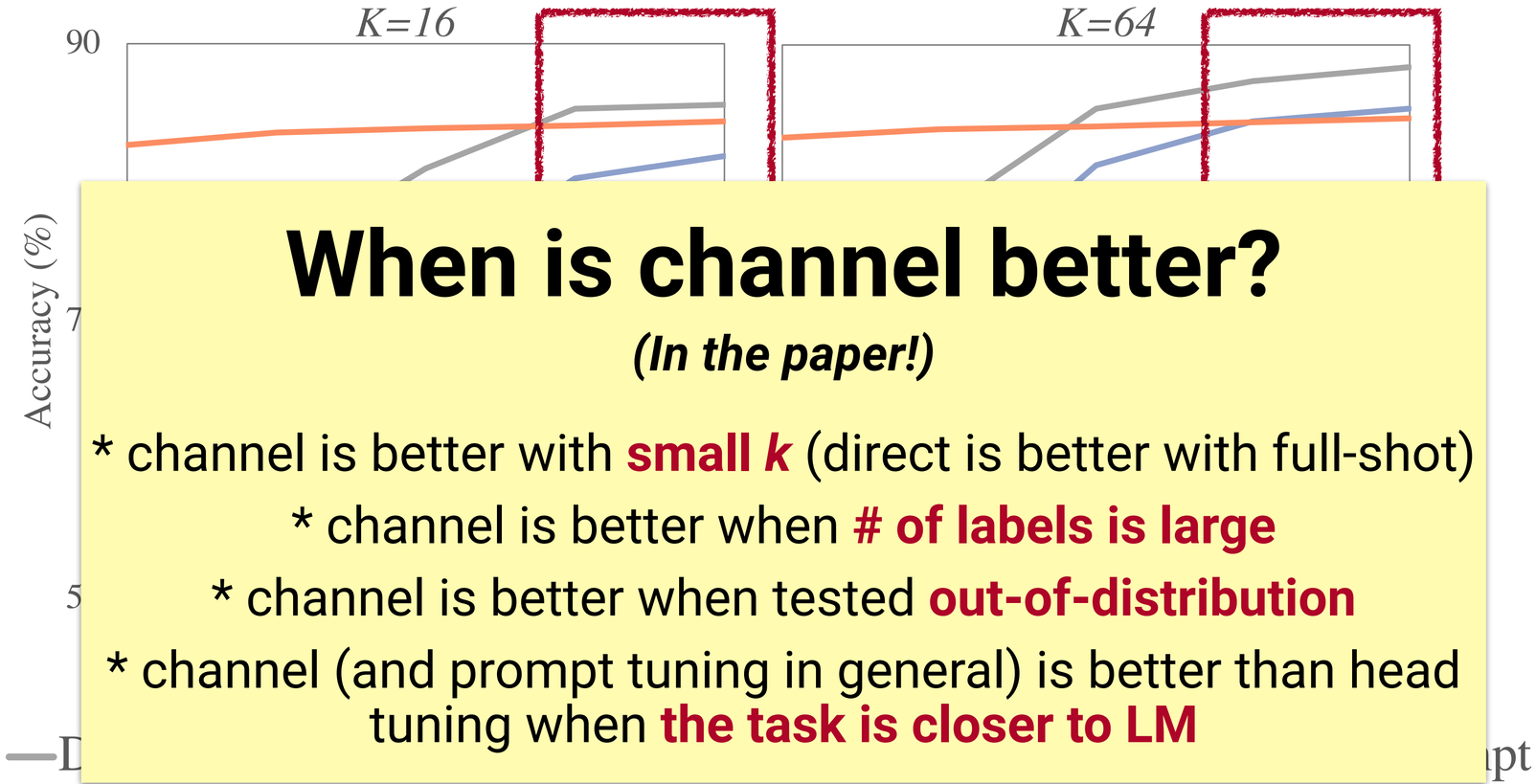
* channel is better when **# of labels is large**

* channel is better when tested **out-of-distribution**

Direct models can be better with perfect balanced data

Ablations

Impact of imbalance in training data



When is channel better?

(In the paper!)

* channel is better with **small *k*** (direct is better with full-shot)

* channel is better when **# of labels is large**

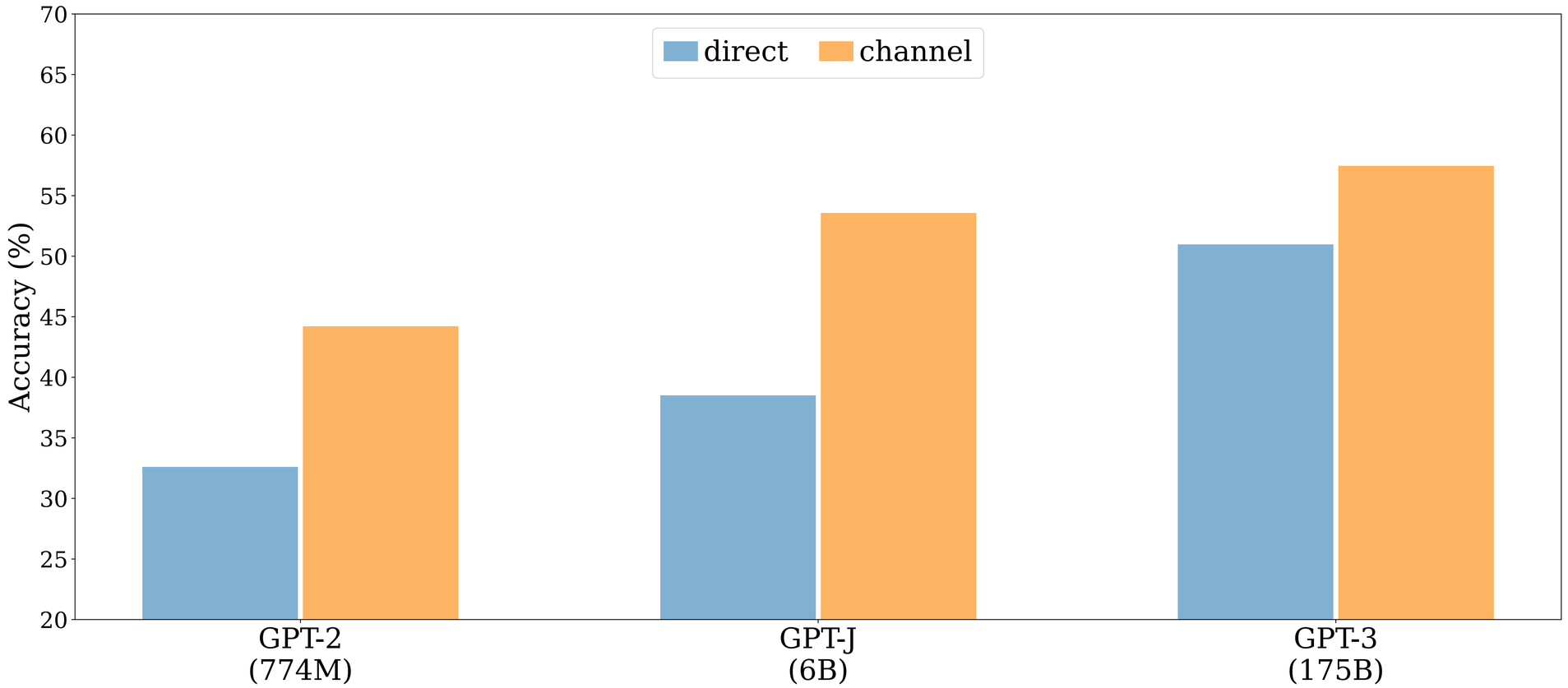
* channel is better when tested **out-of-distribution**

* channel (and prompt tuning in general) is better than head tuning when **the task is closer to LM**

Direct models can be better with perfect balanced data

Extension to larger models

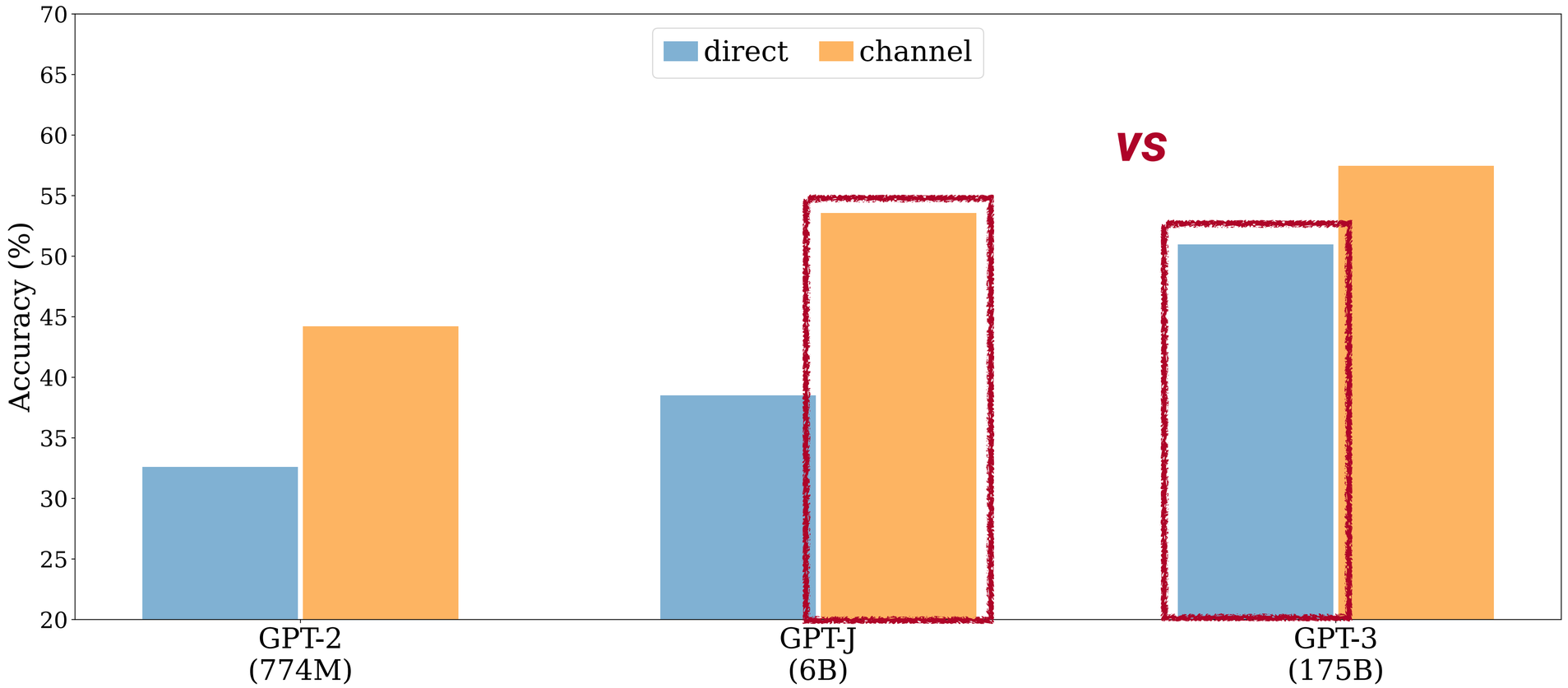
*Channel outperforms direct across **774M to 175B** models*



(on three classification datasets)

Extension to larger models

*Channel outperforms direct across **774M to 175B** models*
*Channel method allows outperforming **x30** bigger models*

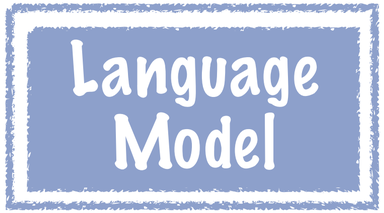


(on three classification datasets)

Takeaways

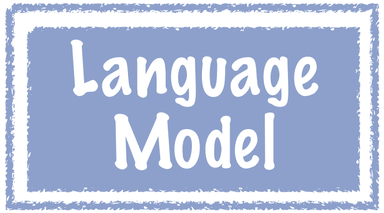
Takeaways

LM Prompting



Takeaways

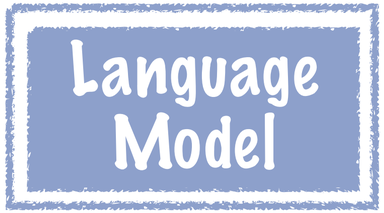
LM Prompting



Variance ↑ Worst-case accuracy ↓

Takeaways

LM Prompting



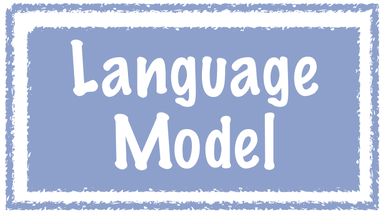
Variance ↑ Worst-case accuracy ↓

Noisy Channel Approach

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)} \propto P(x \mid y)P(y)$$

Takeaways

LM Prompting



Variance ↑ Worst-case accuracy ↓

Noisy Channel Approach

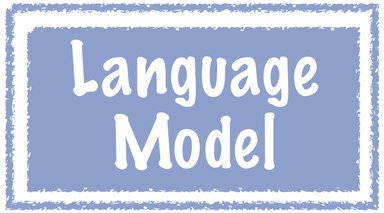
$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)} \propto P(x \mid y)P(y)$$

Variance ↓ Worst-case accuracy ↑

on 10 classification datasets

Takeaways

LM Prompting



Variance ↑ Worst-case accuracy ↓

When is it better?

Noisy Channel Approach

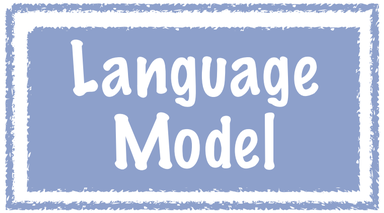
$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)} \propto P(x \mid y)P(y)$$

Variance ↓ Worst-case accuracy ↑

on 10 classification datasets

Takeaways

LM Prompting



Variance ↑ Worst-case accuracy ↓

When is it better?

Small k

Noisy Channel Approach

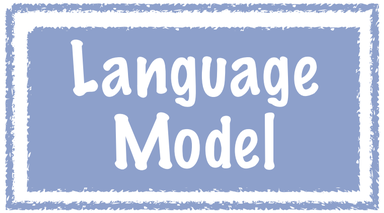
$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)} \propto P(x \mid y)P(y)$$

Variance ↓ Worst-case accuracy ↑

on 10 classification datasets

Takeaways

LM Prompting



Variance ↑ Worst-case accuracy ↓

When is it better?

Small k

Large # of labels

Noisy Channel Approach

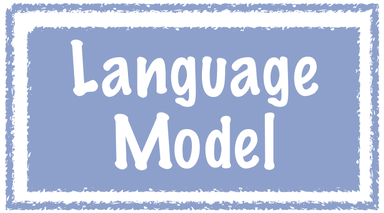
$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)} \propto P(x \mid y)P(y)$$

Variance ↓ Worst-case accuracy ↑

on 10 classification datasets

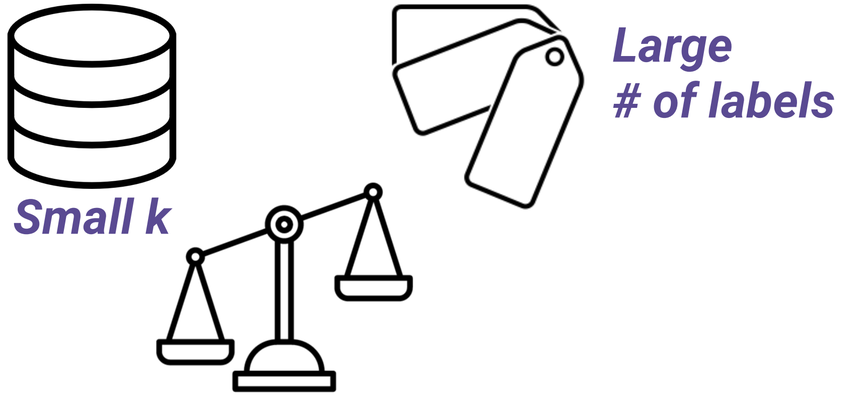
Takeaways

LM Prompting



Variance ↑ Worst-case accuracy ↓

When is it better?



Noisy Channel Approach

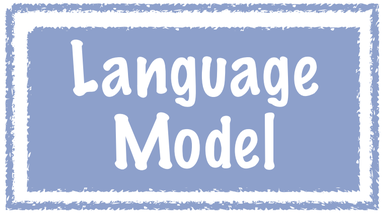
$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)} \propto P(x \mid y)P(y)$$

Variance ↓ Worst-case accuracy ↑

on 10 classification datasets

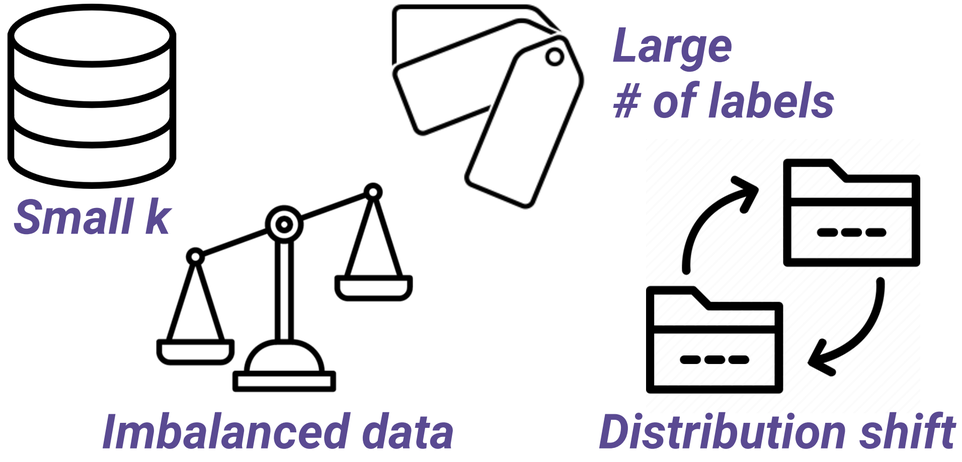
Takeaways

LM Prompting



Variance ↑ Worst-case accuracy ↓

When is it better?



Noisy Channel Approach

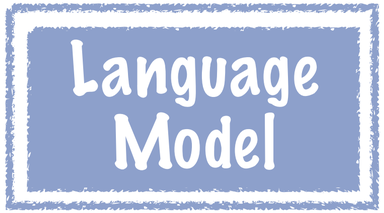
$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)} \propto P(x \mid y)P(y)$$

Variance ↓ Worst-case accuracy ↑

on 10 classification datasets

Takeaways

LM Prompting



Variance ↑ Worst-case accuracy ↓

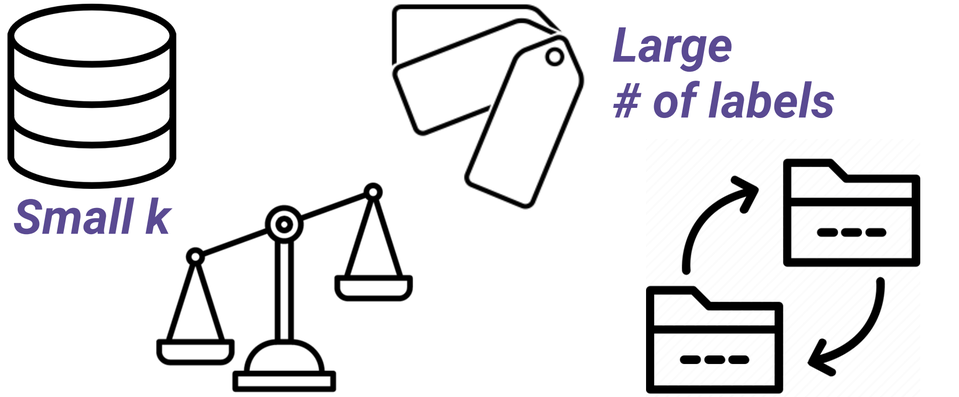
Noisy Channel Approach

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)} \propto P(x \mid y)P(y)$$

Variance ↓ Worst-case accuracy ↑

on 10 classification datasets

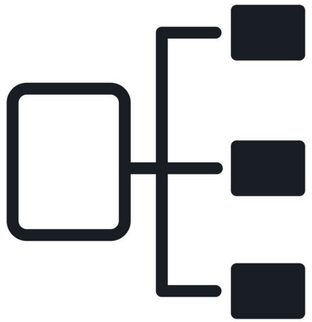
When is it better?



Limitations

Limited to classification

Extension needed!



Thank you for listening

Paper: https://arxiv.org/abs/2108.04106
Code: https://github.com/shmsw25/Channel-LM-Prompting/
Demo: http://qa.cs.washington.edu:2021/

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

UWNLP Meta AI2

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN