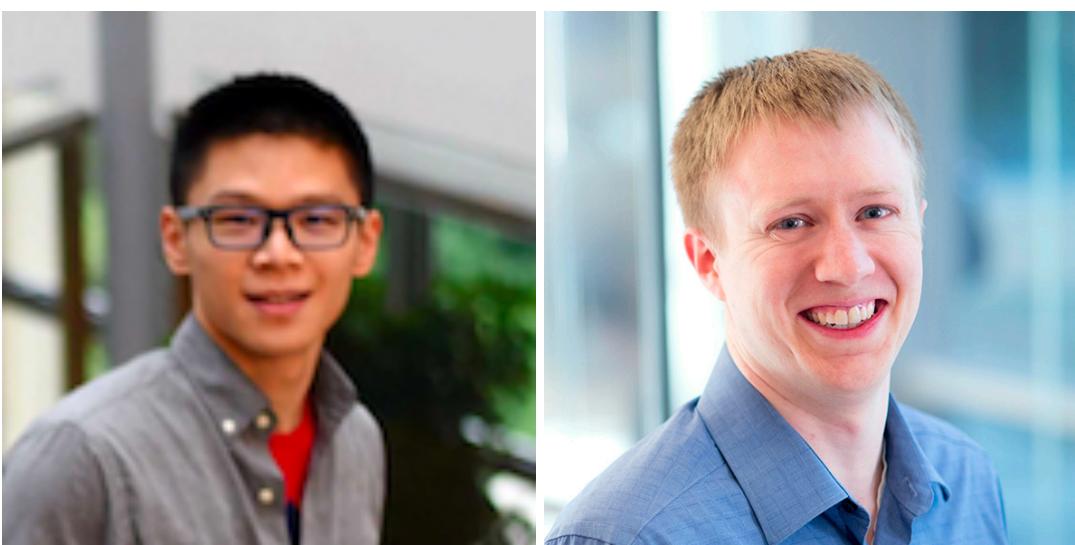


Dissecting Generation Modes for Abstractive Summarization Models via Ablation and Attribution

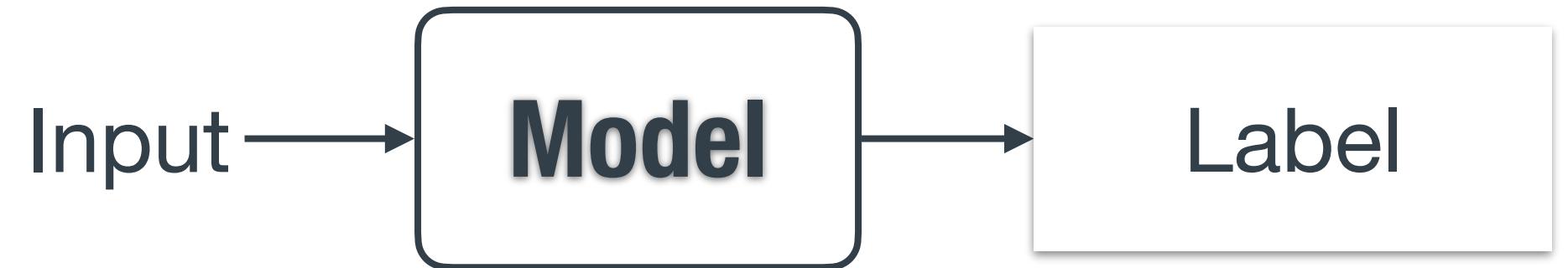
Jiacheng Xu and Greg Durrett

The University of Texas at Austin





Interpreting Summarization Models



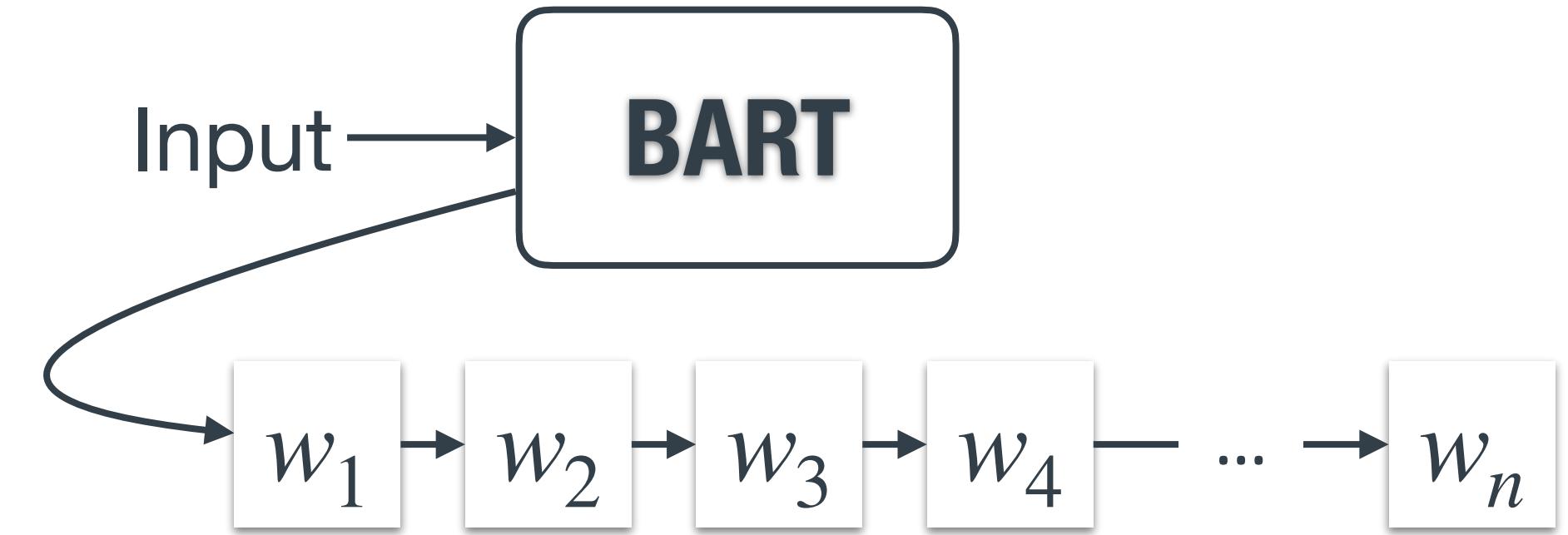
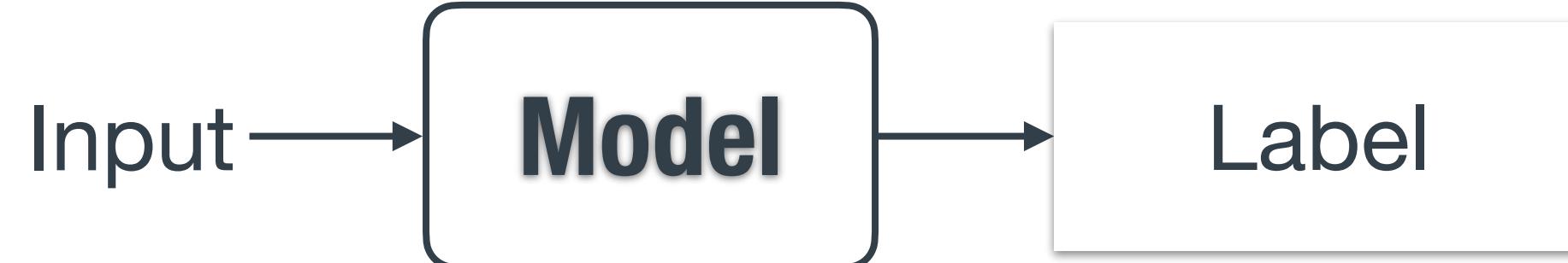
brilliant and moving performances by tom and peter finch
brilliant and moving performances by tom and peter finch
brilliant and moving performances by tom and peter finch

The words "and", "moving", "performances", "by", "tom", "and", "peter", and "finch" are highlighted in different colors across three rows: red, blue, and green.

Well-established techniques for interpreting classification/NLU models



Interpreting Summarization Models



brilliant and moving performances by tom and peter finch
brilliant and moving performances by tom and peter finch
brilliant and moving performances by tom and peter finch

?

Well-established techniques for interpreting classification/NLU models

How to interpret complicated sequential decisions?



Overview

Our contribution: a two-stage decision interpretation framework for summarization



For each time step: (1) Does the model need input context?
(2) If yes, which input tokens matter?



Overview

Our contribution: a two-stage decision interpretation framework for summarization



For each time step: (1) Does the model need input context?
(2) If yes, which input tokens matter?

Ablation: what if a more basic model can do the job?

- in 20% of cases our model functions as a LM



Overview

Our contribution: a two-stage decision interpretation framework for summarization



For each time step: (1) Does the model need input context?
(2) If yes, which input tokens matter?

Ablation: what if a more basic model can do the job?

- in 20% of cases our model functions as a LM

Attribution: what part of the input leads to the decision?

- many attribution methods; what works best?
- we propose a framework for evaluation



Overview

Our contribution: a two-stage decision interpretation framework for summarization



For each time step: (1) Does the model need input context?
(2) If yes, which input tokens matter?

Ablation: what if a more basic model can do the job?

- in 20% of cases our model functions as a LM

Attribution: what part of the input leads to the decision?

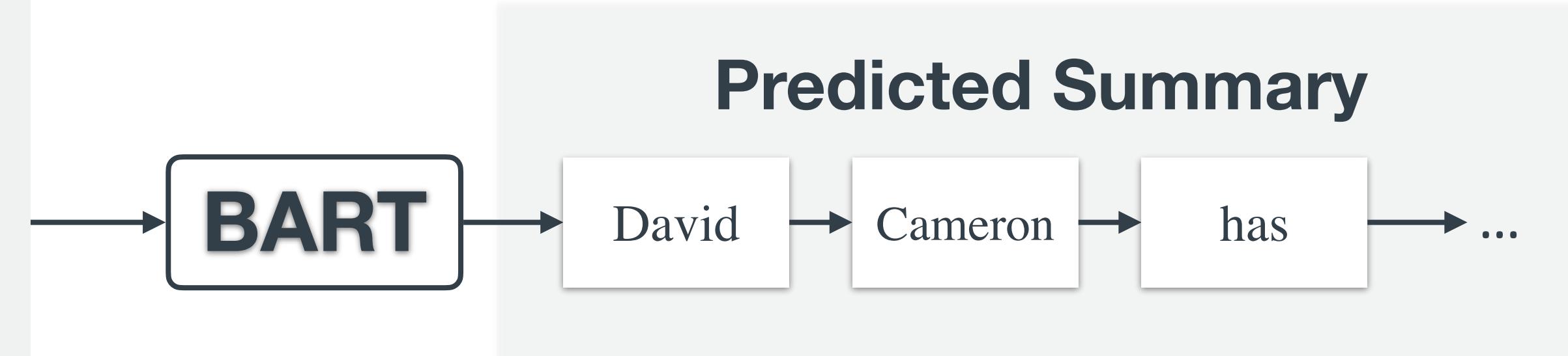
- many attribution methods; what works best?
- we propose a framework for evaluation



Setup

Input Article

Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival [...]



Predicted Summary

BART

David

→ Cameron

has

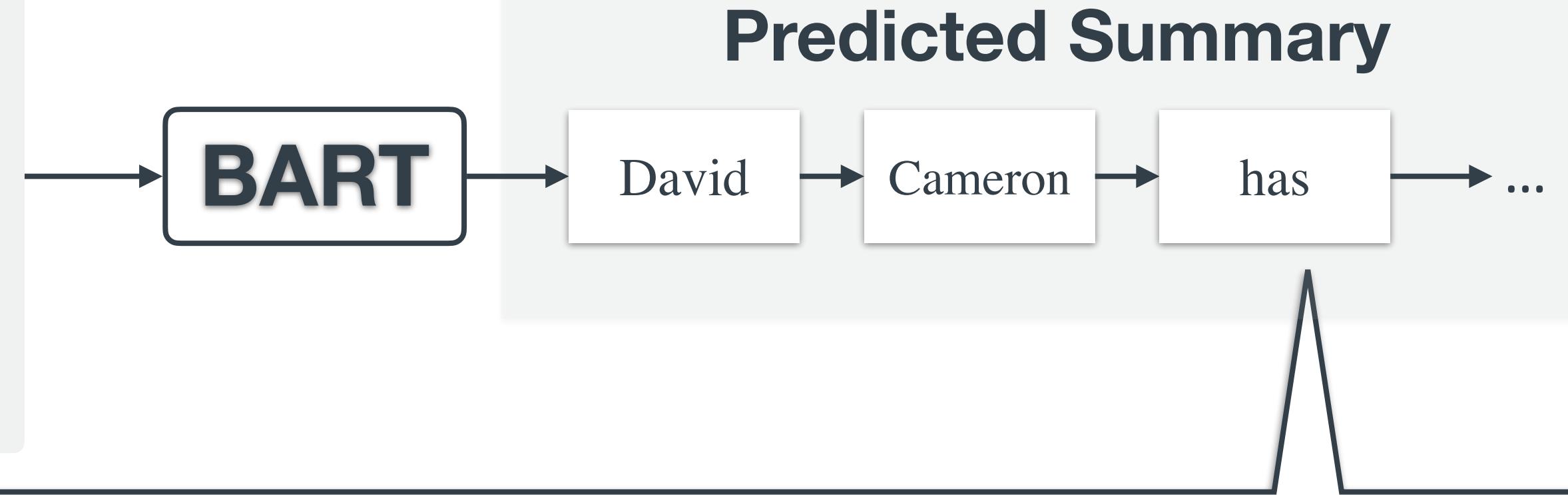
→ ...



Setup

Input Article

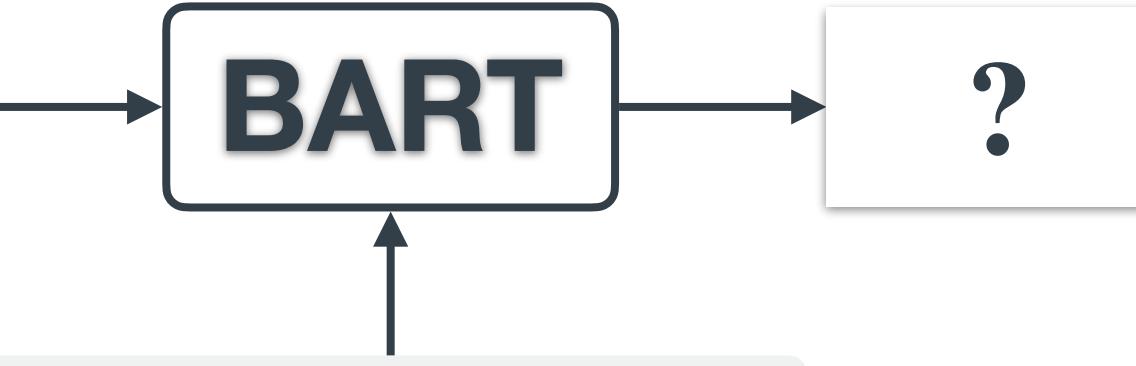
Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival [...]



Input Article

Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister [...]

For each time step, we provide input and prefix, and the model predicts the next token.



Prefix

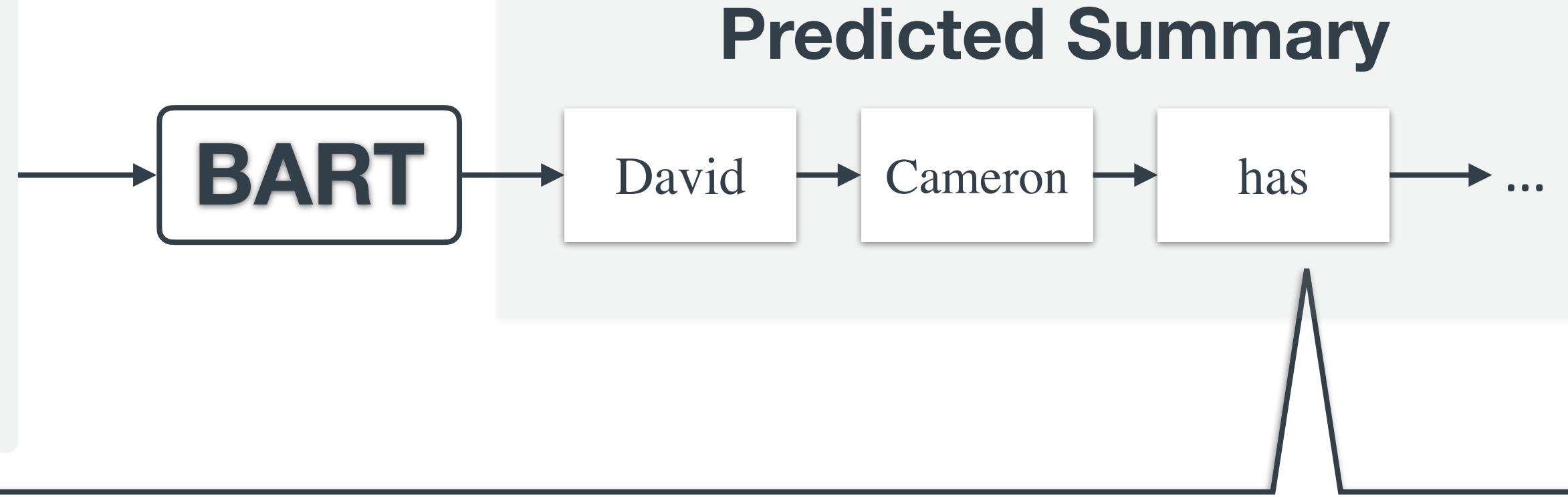
David Cameron



Setup

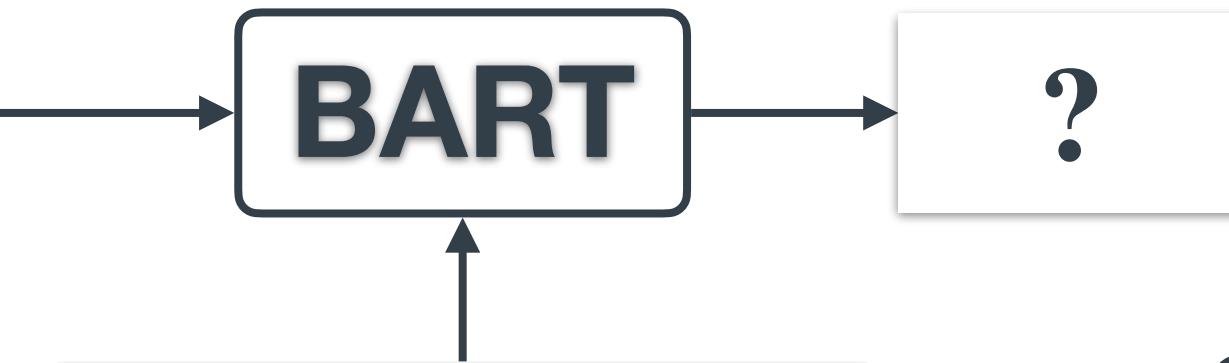
Input Article

Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival [...]



Empty or a subset of tokens

For each time step, we provide input and prefix, and the model predicts the next token.



Prefix
David Cameron

Always provided
in all settings!



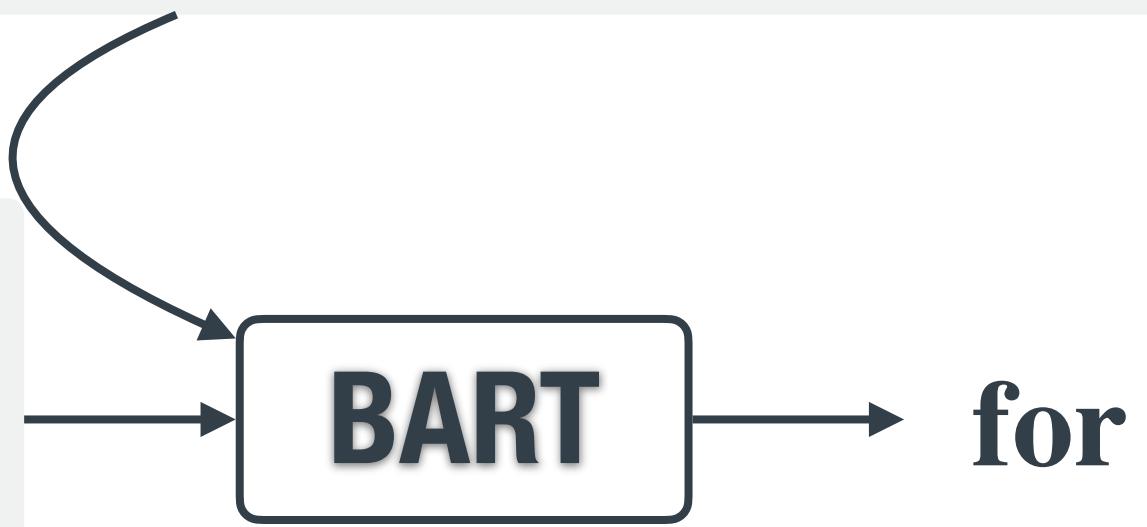
Ablation

Input Article

Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival [...]

Prefix

David Cameron has urged Londoners to vote





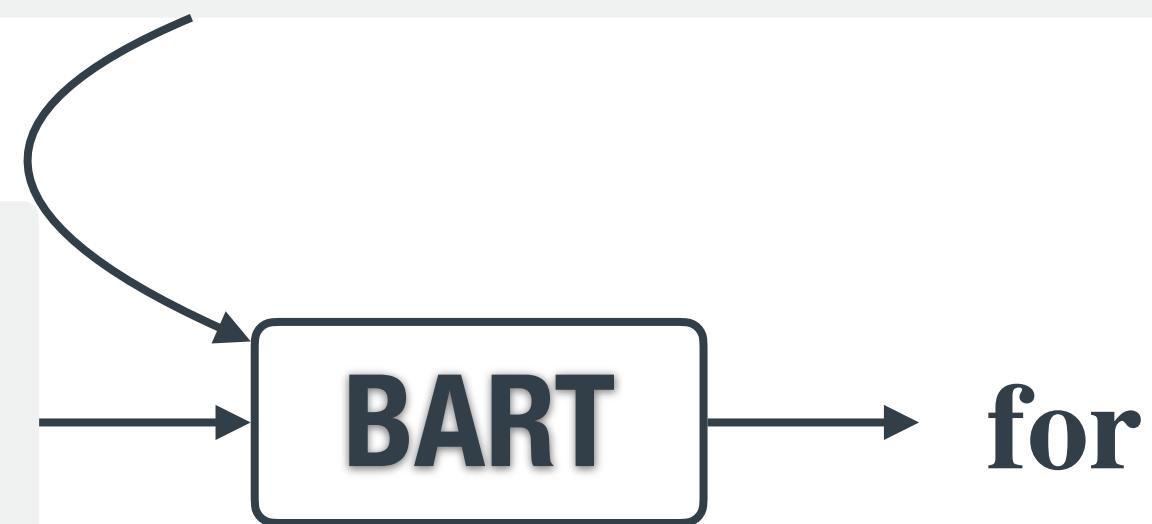
Ablation

Input Article

Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival [...]

Prefix

David Cameron has urged Londoners to vote



Human: Why does the model say “**for**”?



Model: I am confident! Do my *ablated* versions agree with me?



Model

$$p(\text{ for} \mid \text{graduation cap icon}, \text{document icon}, \text{prefix}) = 0.96$$



Agree!



Model
w/o Input Article

$$p(\text{ for} \mid \text{graduation cap icon}, \text{crossed-out document icon}, \text{prefix}) = 0.95$$

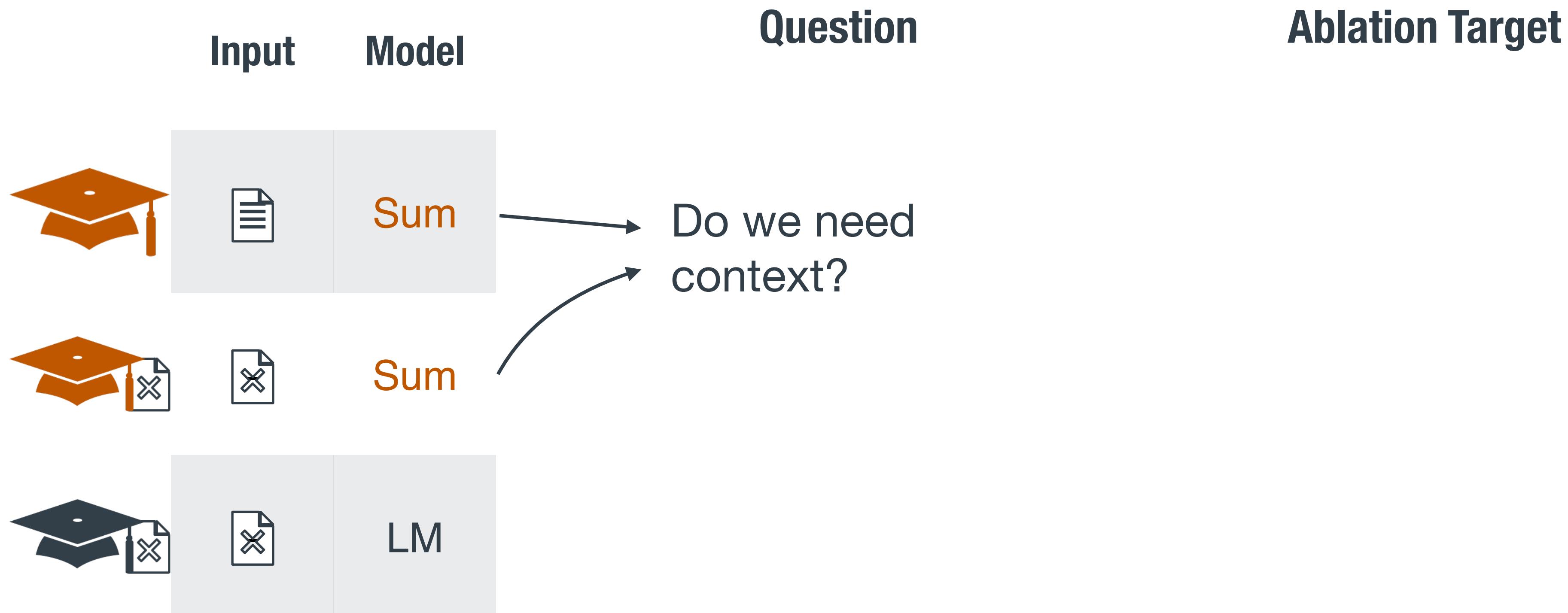


Ablation of

Input	Model	Question	Ablation Target
		Sum	
		Sum	
		LM	

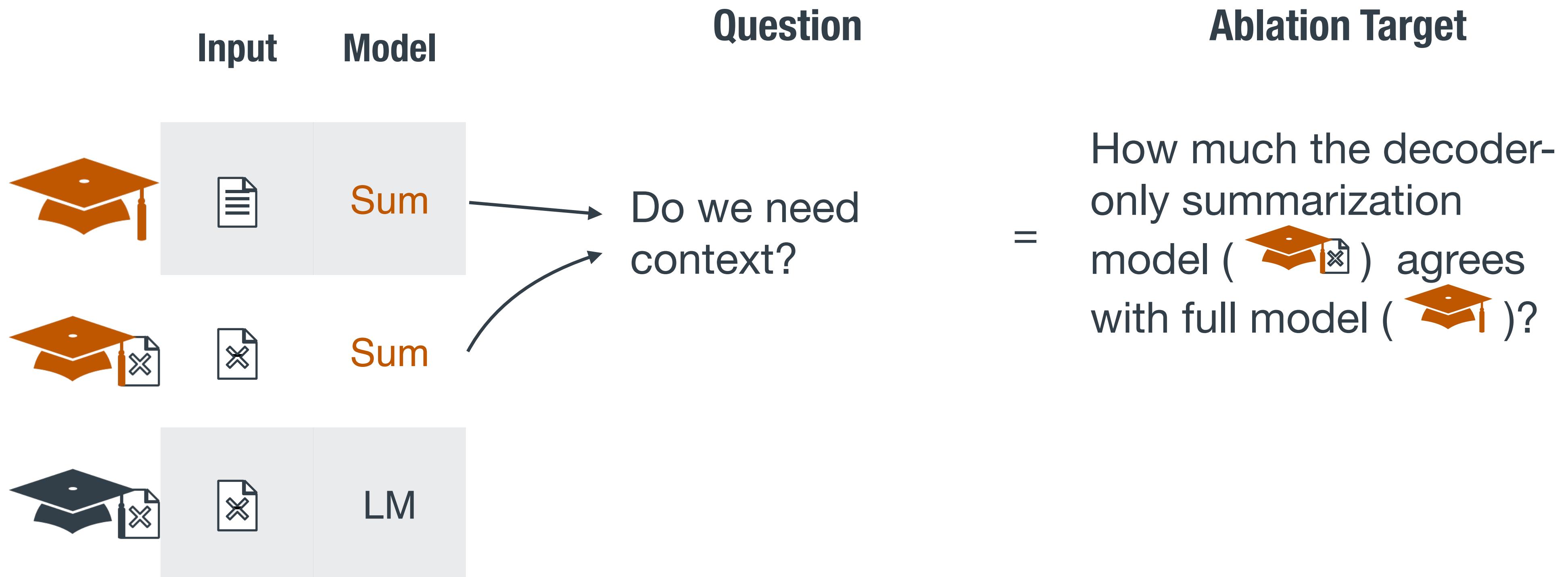


Ablation of



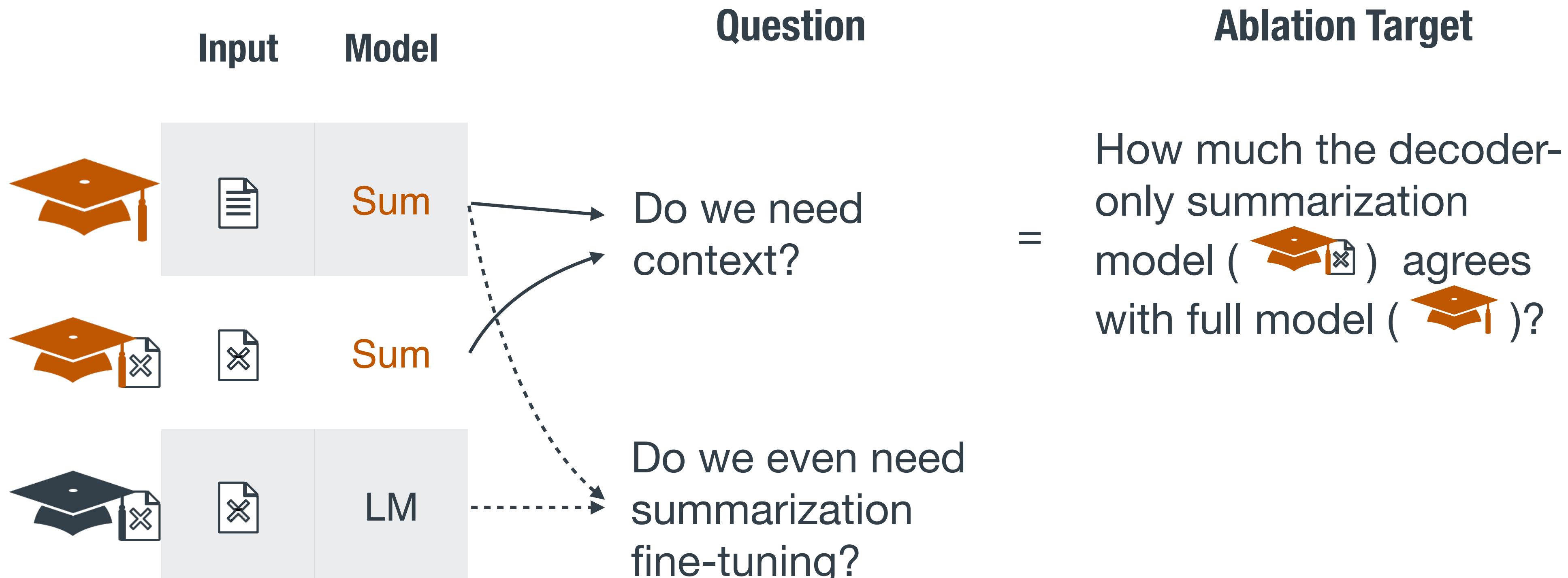


Ablation of



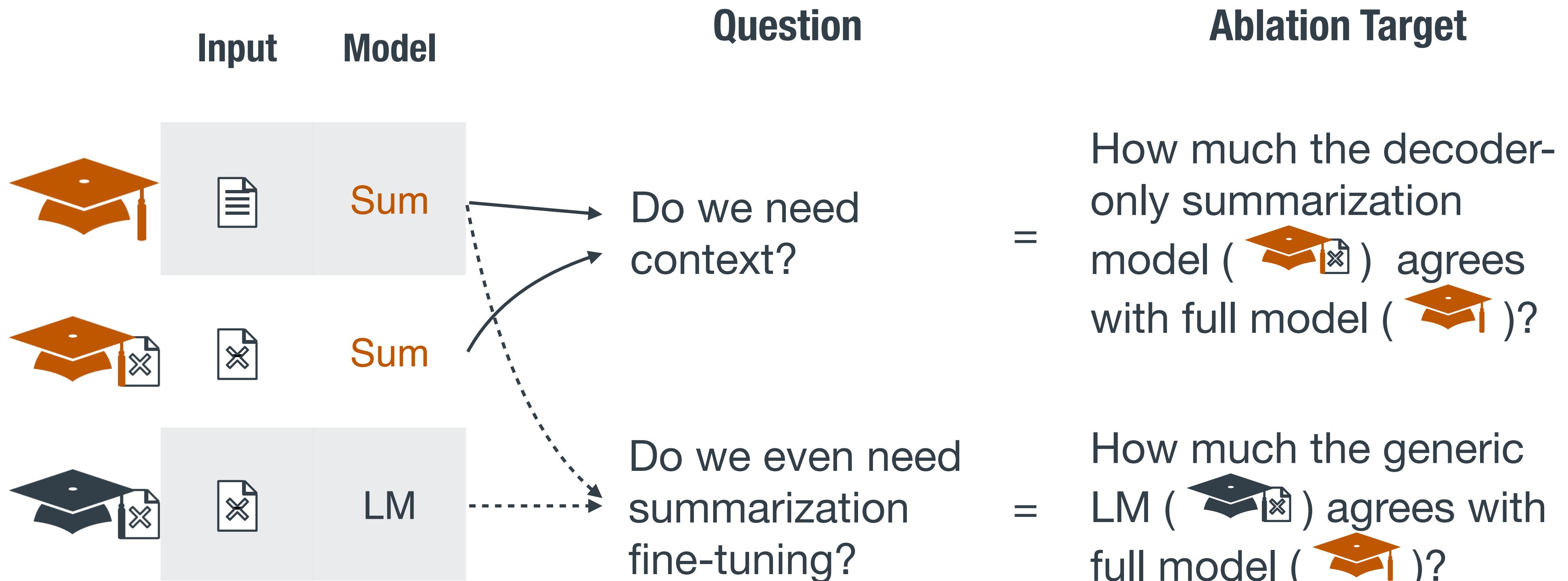


Ablation of





Ablation of





Ablation of

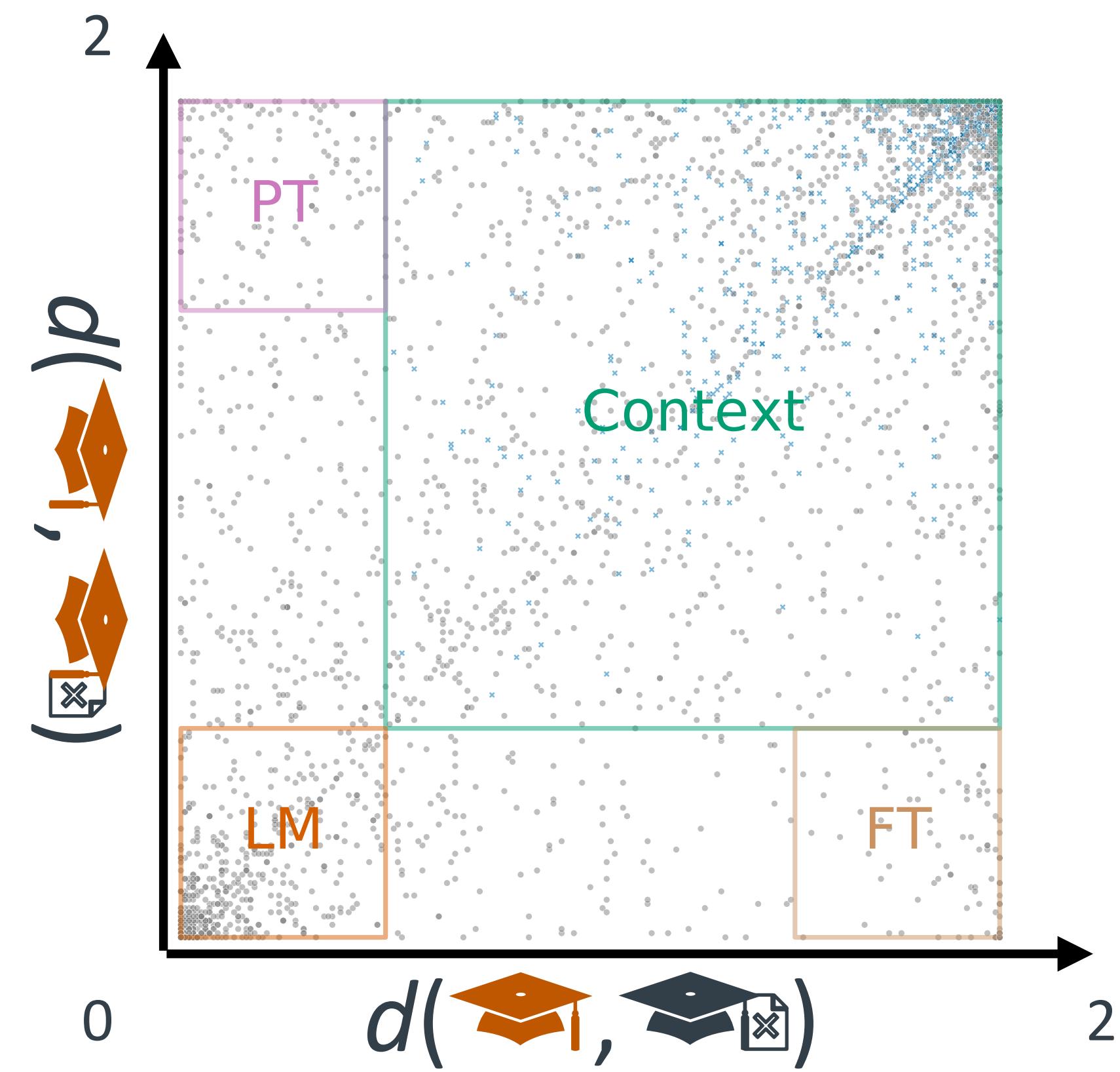
Input	Model	Question	Ablation Target	Axis on Map
	Sum	Do we need context?	= How much the decoder-only summarization model () agrees with full model ()?	y
	Sum			
	LM	Do we even need summarization fine-tuning?	= How much the generic LM () agrees with full model ()?	x



Mapping Generation Modes

Distance function

$$d(p, q) = \sum_i^{|V|} |p_i - q_i|$$

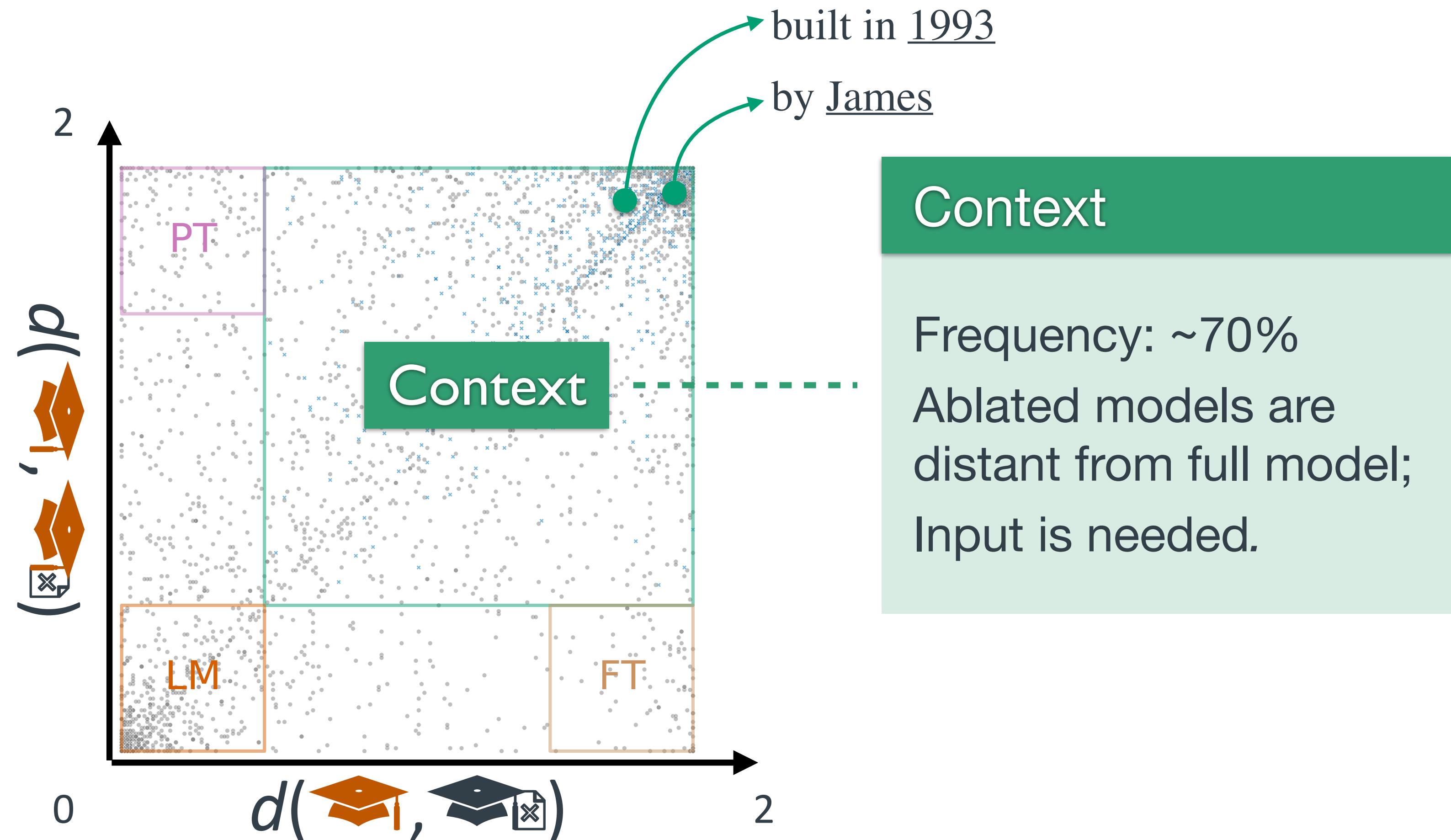




Mapping Generation Modes

Distance function

$$d(p, q) = \sum_i^{|V|} |p_i - q_i|$$

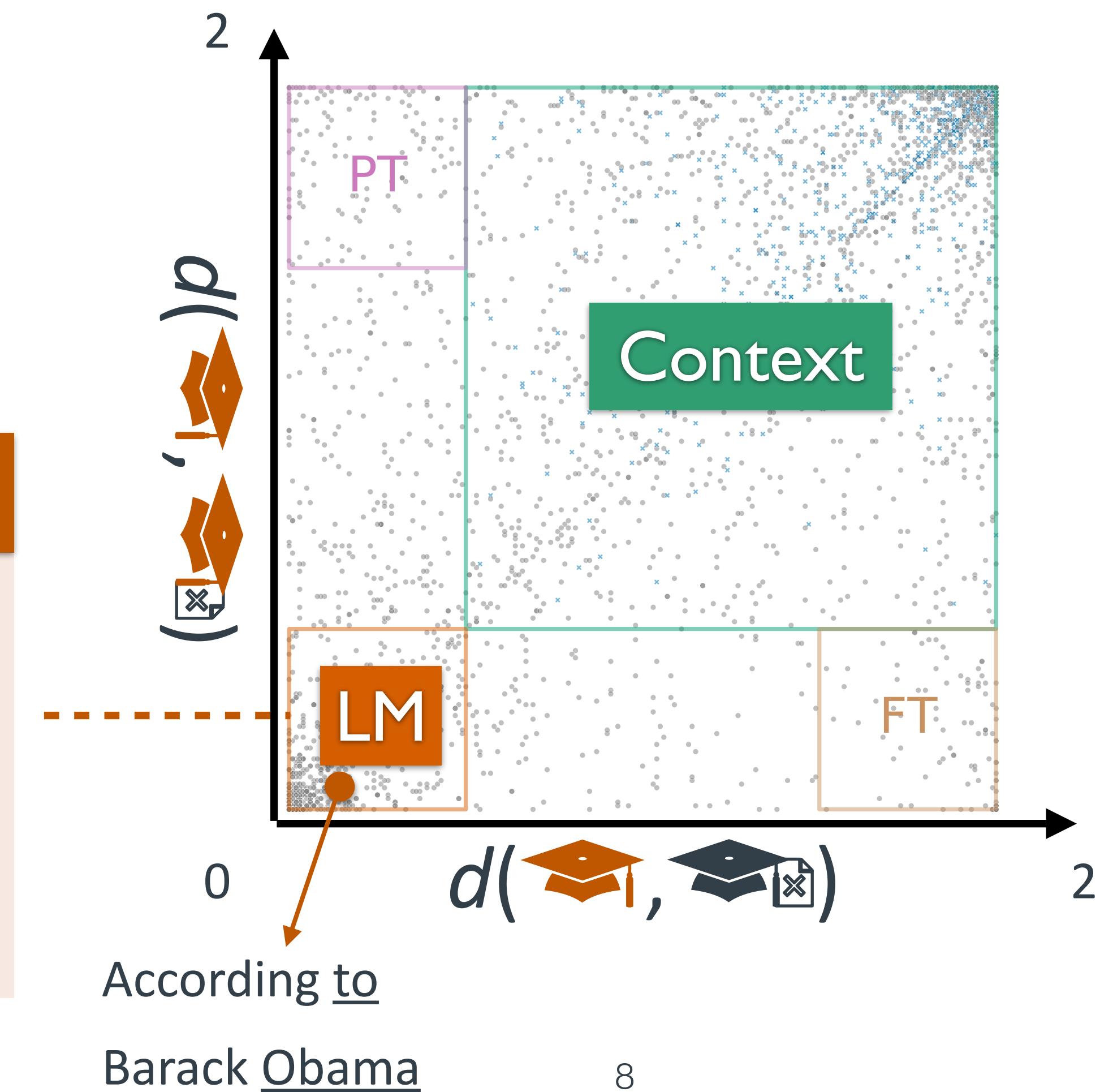




Mapping Generation Modes

LM

Frequency: ~20%
All ablated models
agree with full model;
Input is *not* needed.

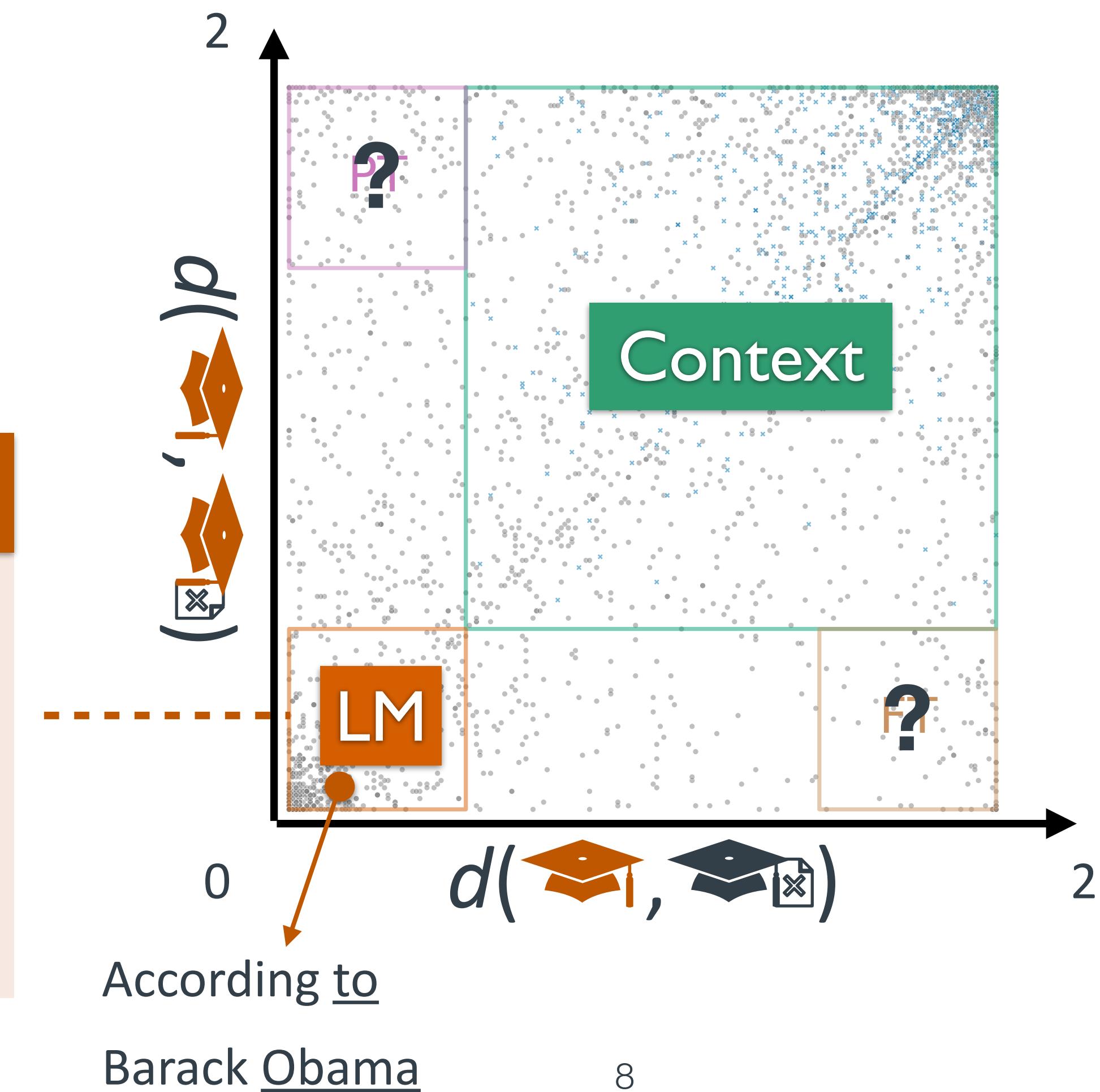




Mapping Generation Modes

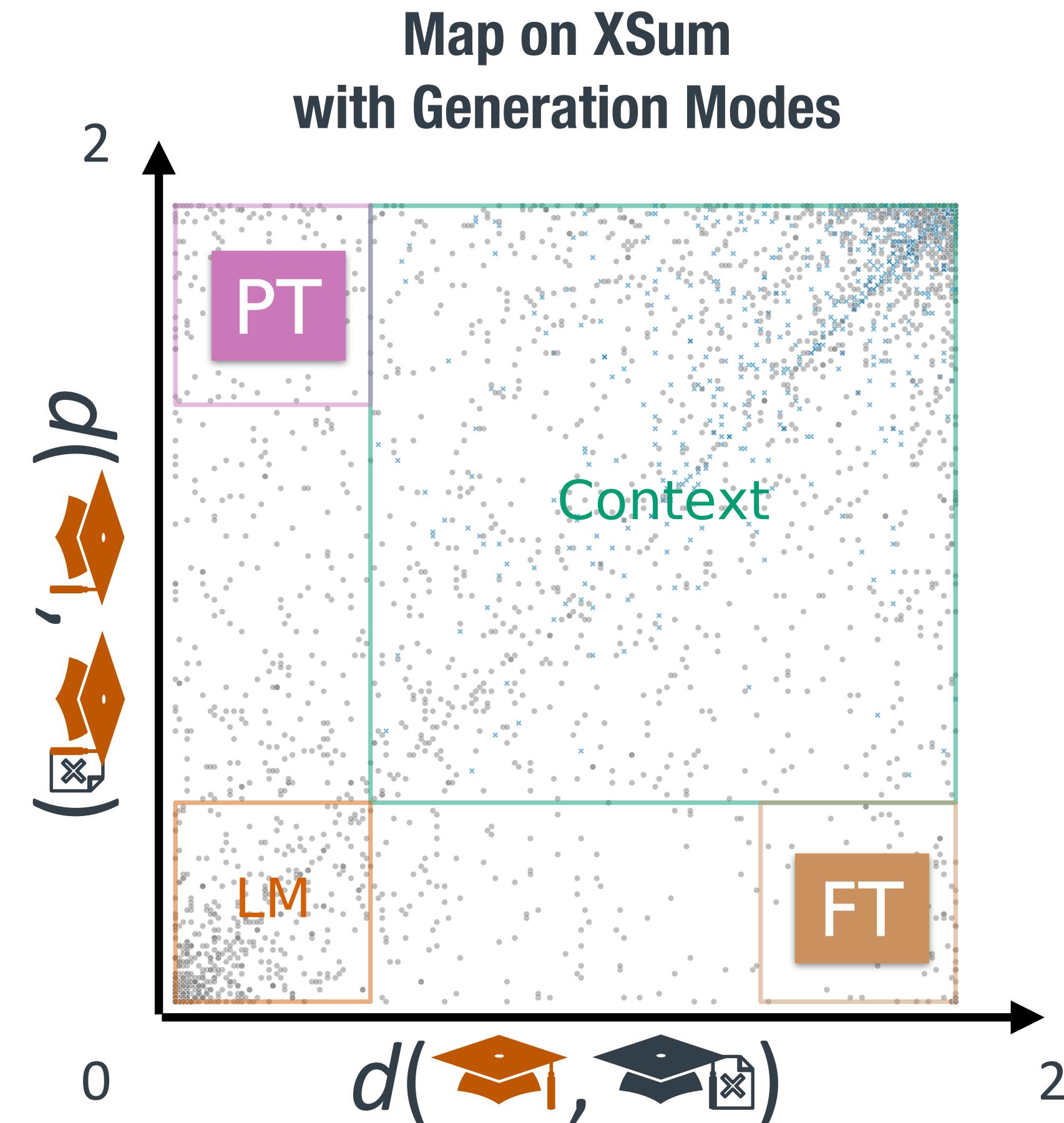
LM

Frequency: ~20%
All ablated models
agree with full model;
Input is *not* needed.





Memorization Bias





Memorization Bias

Pre-Training bias

Frequency: ~2%

Fine-tuning a LM on training data causes it to work less well.

Example

Gail Scott was desperate to emulate Kylie __?__



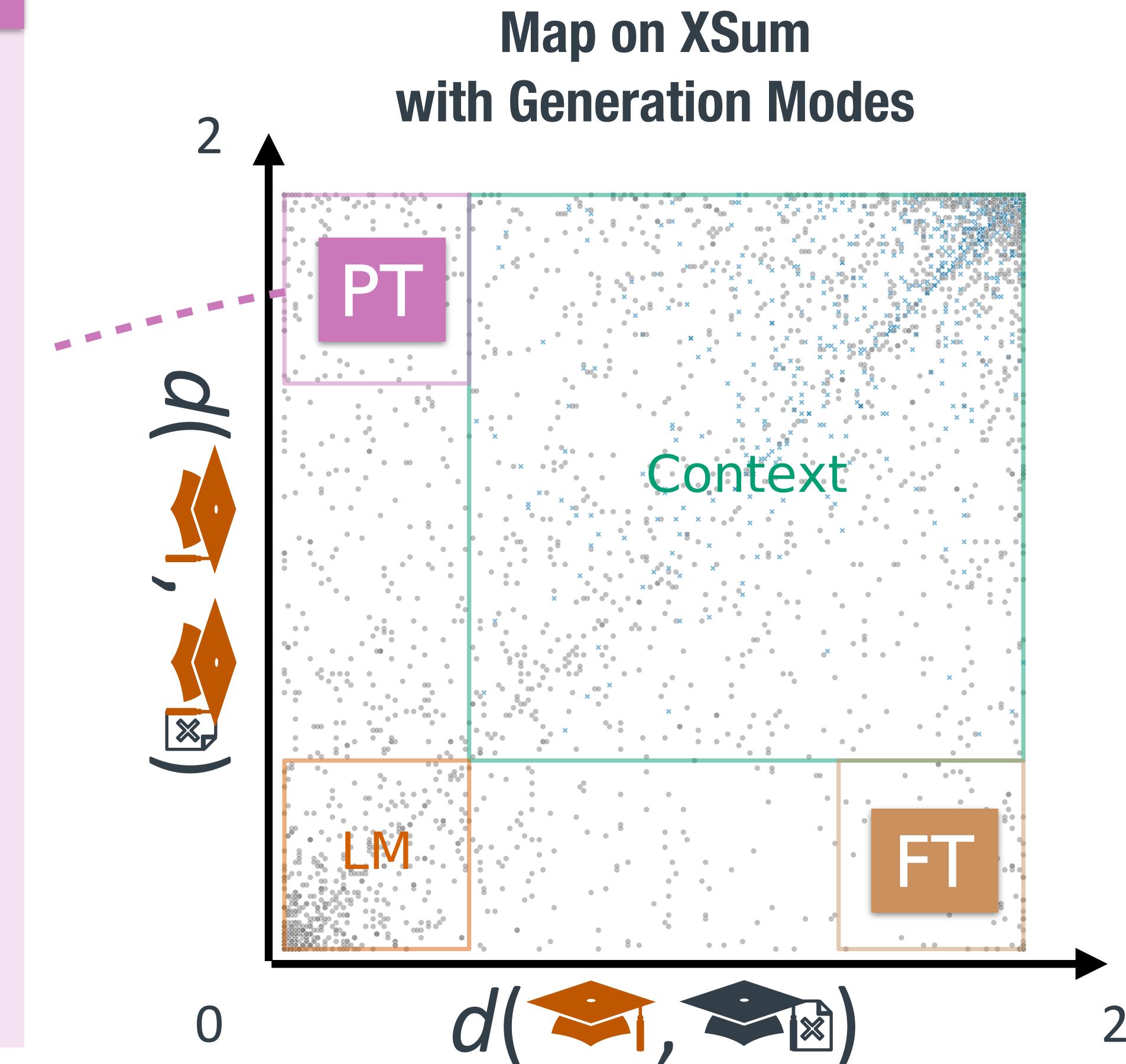
$$p(\text{Jenner}) = 0.90$$



$$p(\text{Minogue}) = 0.80$$



$$p(\text{Jenner}) = 0.99$$





Memorization Bias

Pre-Training bias

Frequency: ~2%

Fine-tuning a LM on training data causes it to work less well.

Example

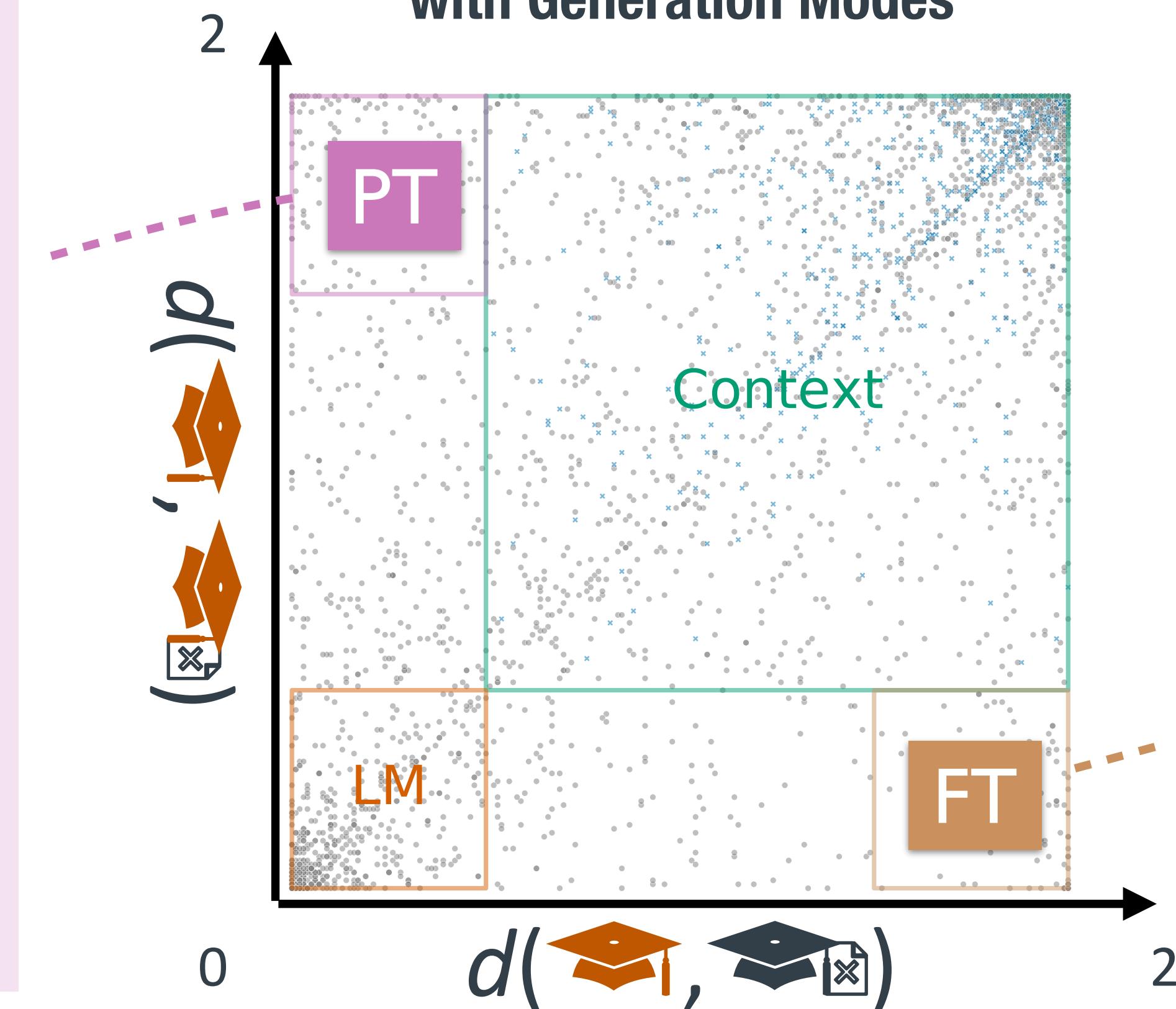
Gail Scott was desperate to emulate Kylie __?__

$p(\text{Jenner}) = 0.90$

$p(\text{Minogue}) = 0.80$

$p(\text{Jenner}) = 0.99$

Map on XSum
with Generation Modes



Fine-Tuning bias

Frequency: ~2%

Fine-tuned decoder-only model without input is a close match but the pre-trained LM is not.

Example

In our series of letters from African journalists, [...]

0.5% of ref summaries in XSum



Memorization Bias

Pre-Training bias

Frequency: ~2%

Fine-tuning a LM on training data causes it to work less well.

Example

Gail Scott was desperate to emulate Kylie __?__



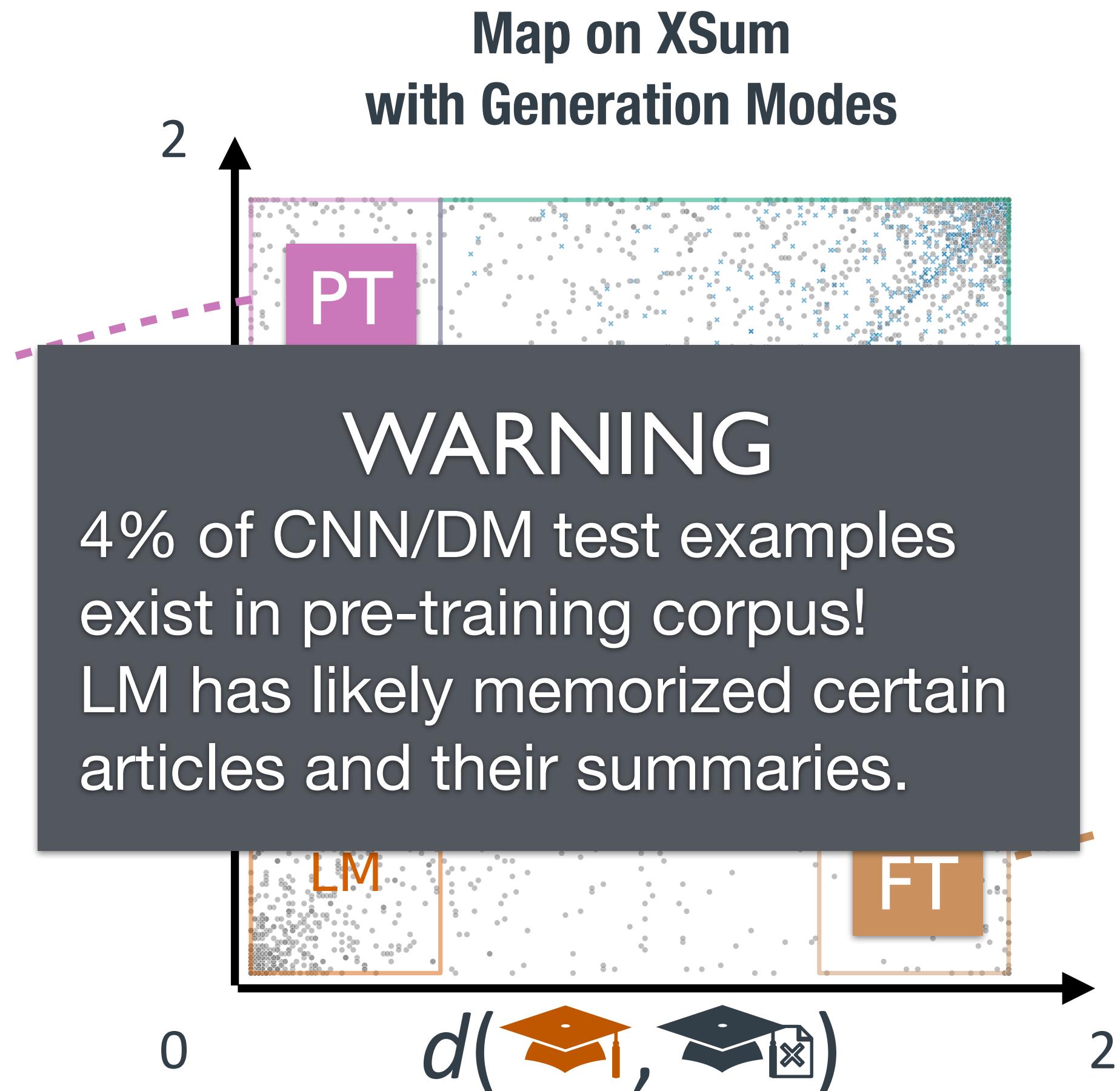
$$p(\text{Jenner}) = 0.90$$



$$p(\text{Minogue}) = 0.80$$



$$p(\text{Jenner}) = 0.99$$



Fine-Tuning bias

Frequency: ~2%

Fine-tuned decoder-only model without input is a close match but the pre-trained LM is not.

Example

In our series of letters from African journalists, [...]

0.5% of ref summaries in XSum



Overview

Our contribution: a two-stage decision interpretation framework for summarization



For each time step: (1) Does the model need input context?
(2) If yes, which input tokens matter?

Ablation: what if a more basic model can do the job?

- in 20% of cases our model functions as a LM

Attribution: what part of the input leads to the decision?

- many attribution methods; what works best?
- we propose a framework for evaluation

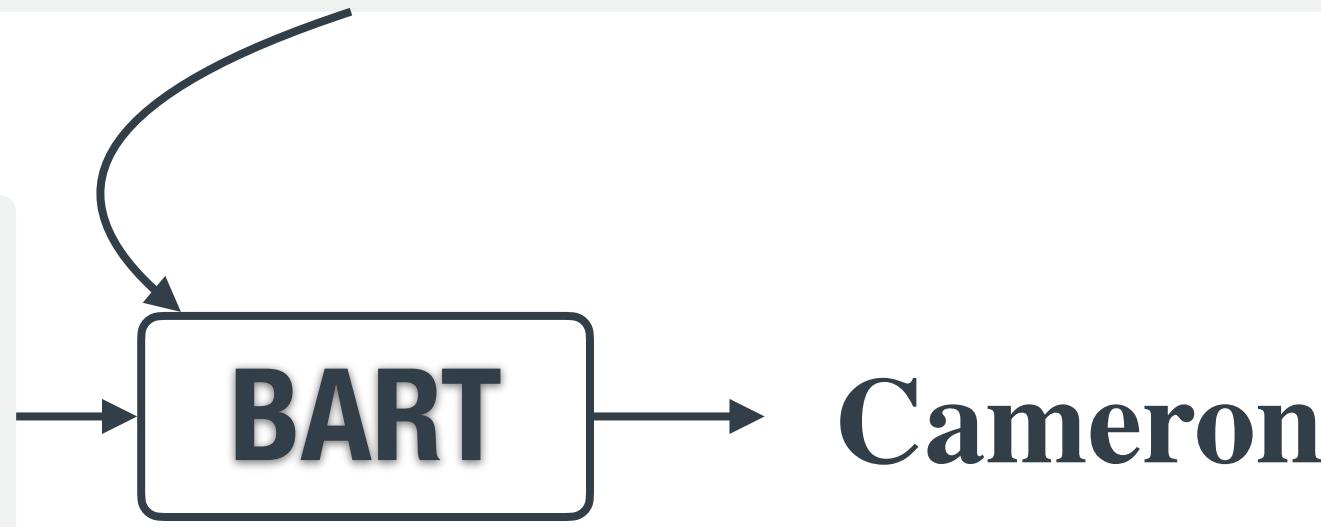


Attribution

Input Article

Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival [...]

Prefix
David

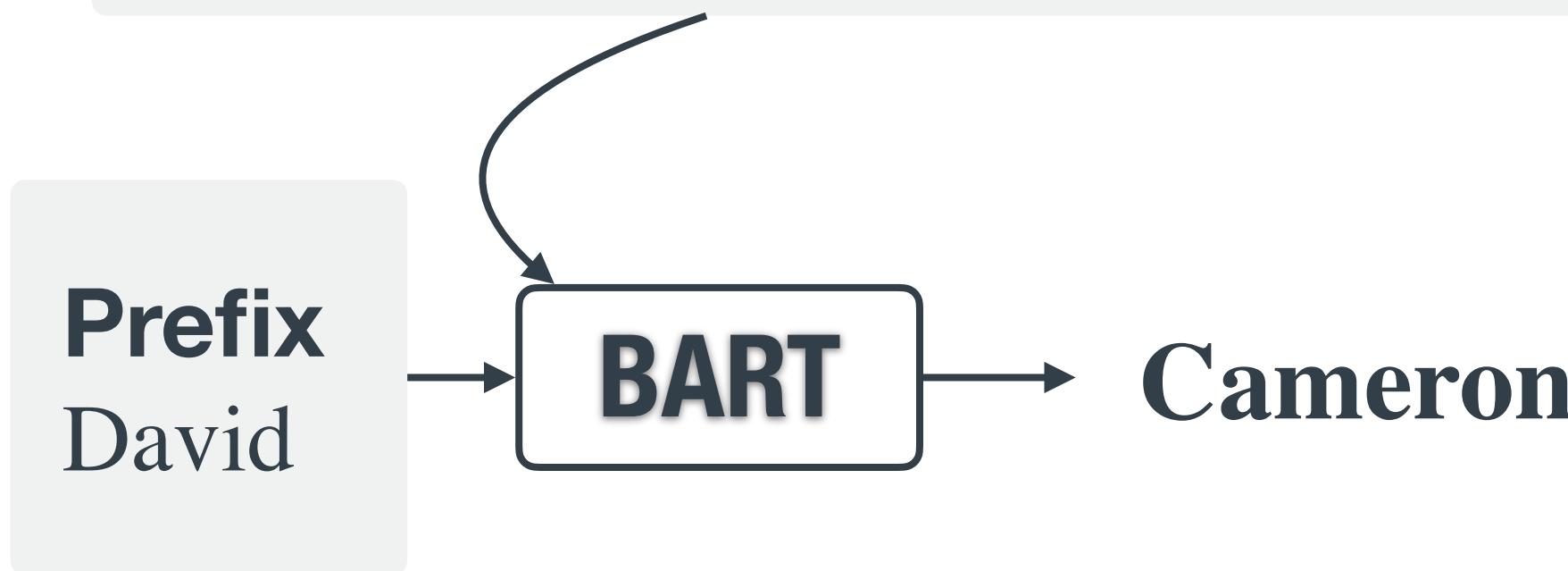




Attribution

Input Article

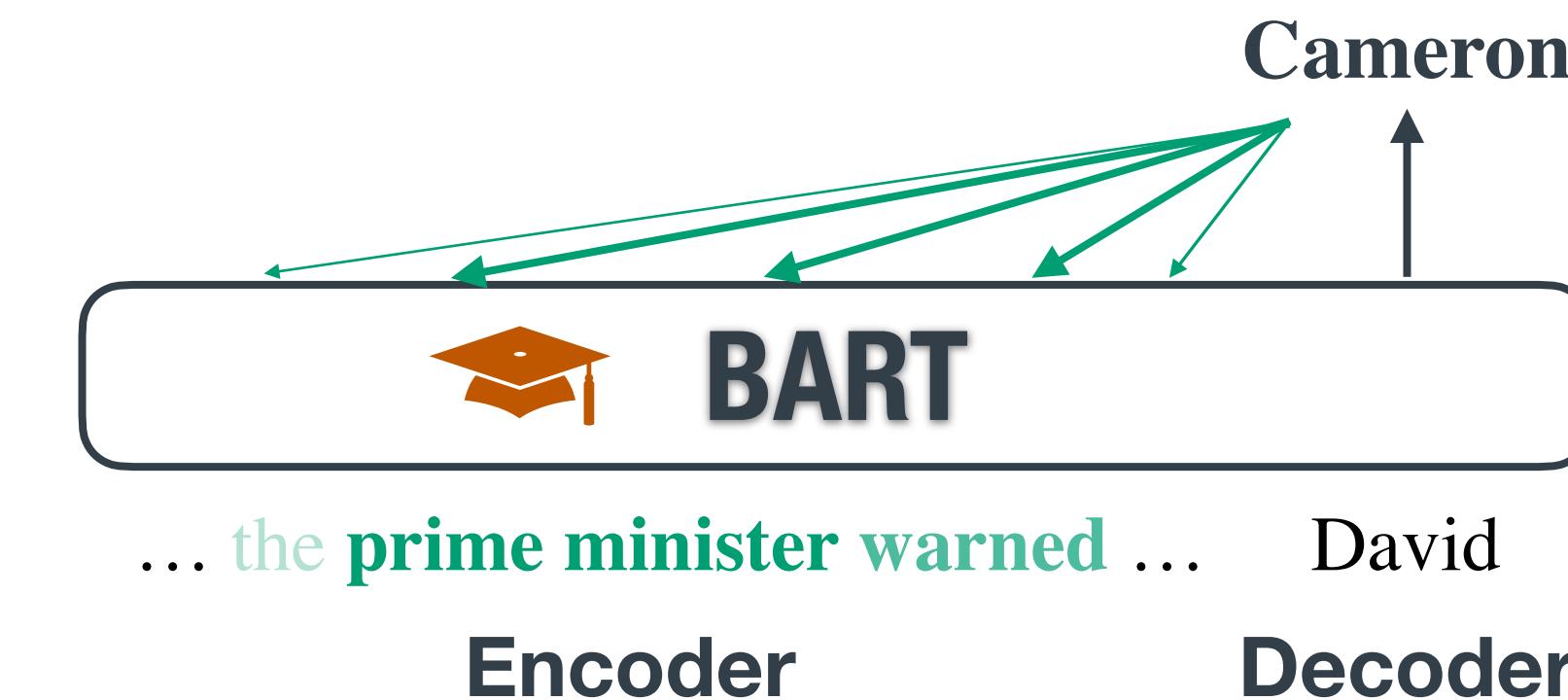
Speaking at a rally for Tory candidate Zac Goldsmith, the prime minister warned about the dangers of a Labour victory for the capital's economy. Mr Goldsmith said his Labour rival [...]



- Human: Why does the model say “**Cameron**”?
- Model: Ablated version disagrees → Input matters.
- Human: So what exactly do you look at?

Find the context which actually matters!

Methods: gradient, occlusion, etc.





Attribution

#s The country 's consumer watchdog has taken Apple to court for false advertising because the tablet computer does not work on Australia 's 4 G network . Apple 's lawyers said they were willing to publish a clarification . However the company does not accept that it misled customers . The Australian Competition and Consumer Commission (AC CC) said on Tuesday

Integrated Gradient

#s The country 's consumer watchdog has taken Apple to court for false advertising because the tablet computer does not work on Australia 's 4 G network . Apple 's lawyers said they were willing to publish a clarification . However the company does not accept that it misled customers . The Australian Competition and Consumer Commission (AC CC) said on Tuesday

Attention

Challenge

- hard to compare highlights
- inconsistency among different attribution methods

What we want

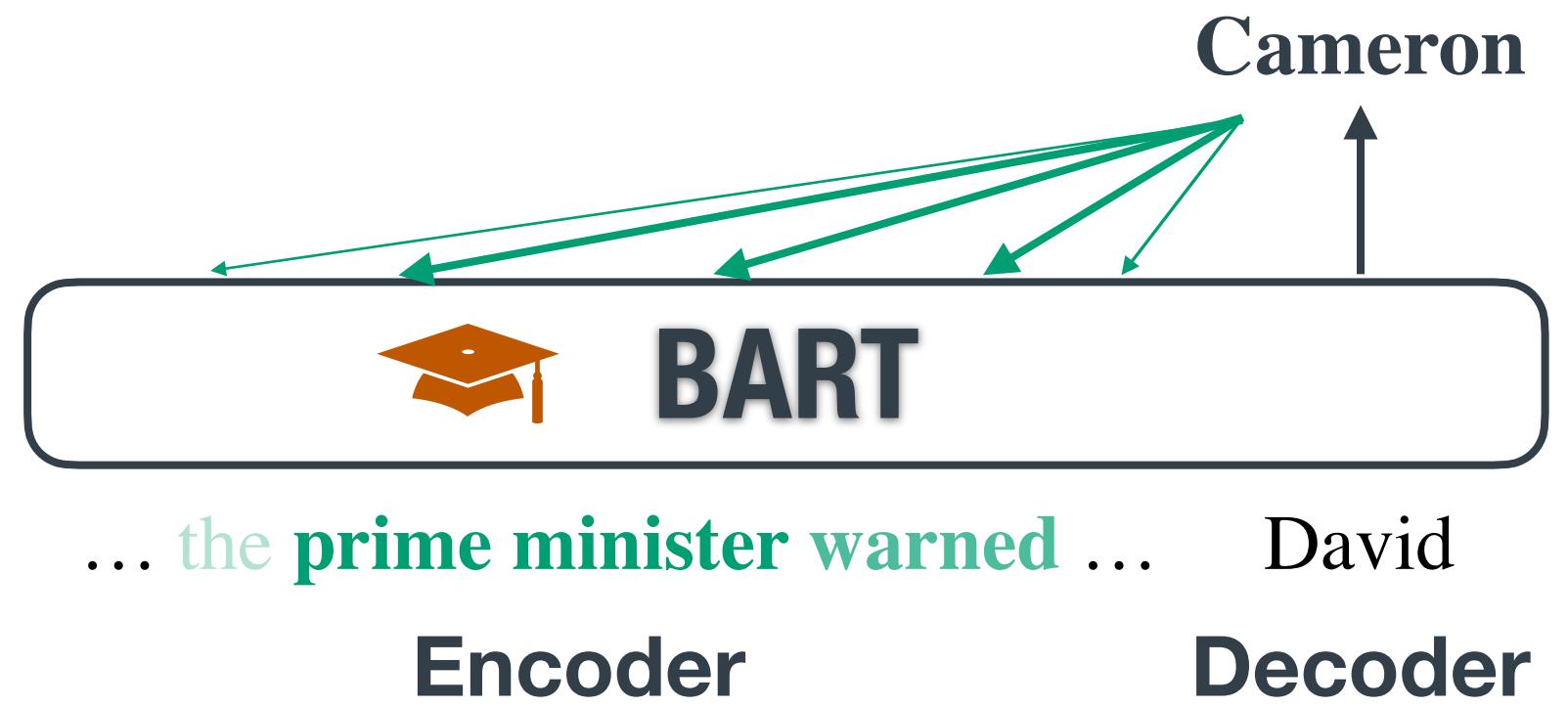
an evaluation protocol for attributions



Evaluation Protocol



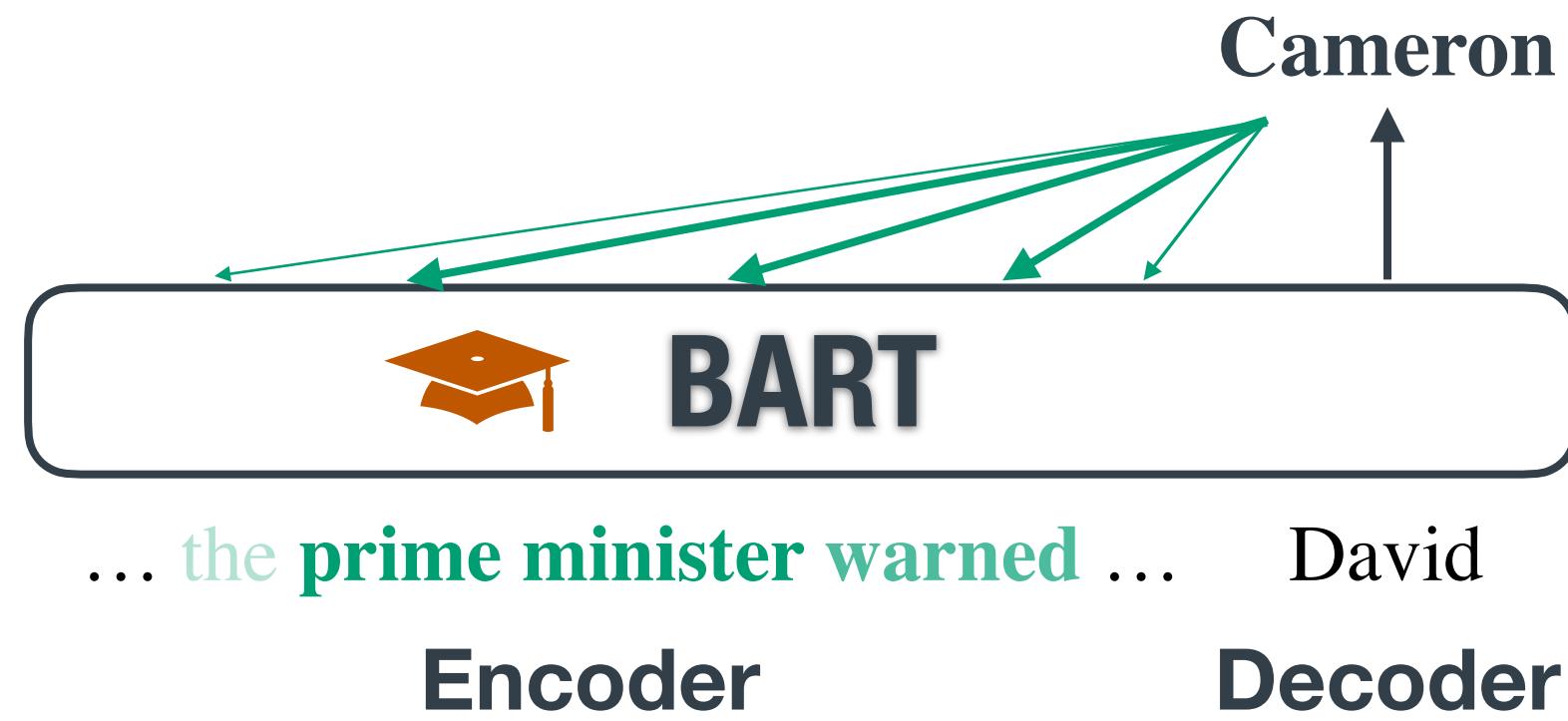
Evaluation Protocol



Assign score for each token
according to attribution



Evaluation Protocol



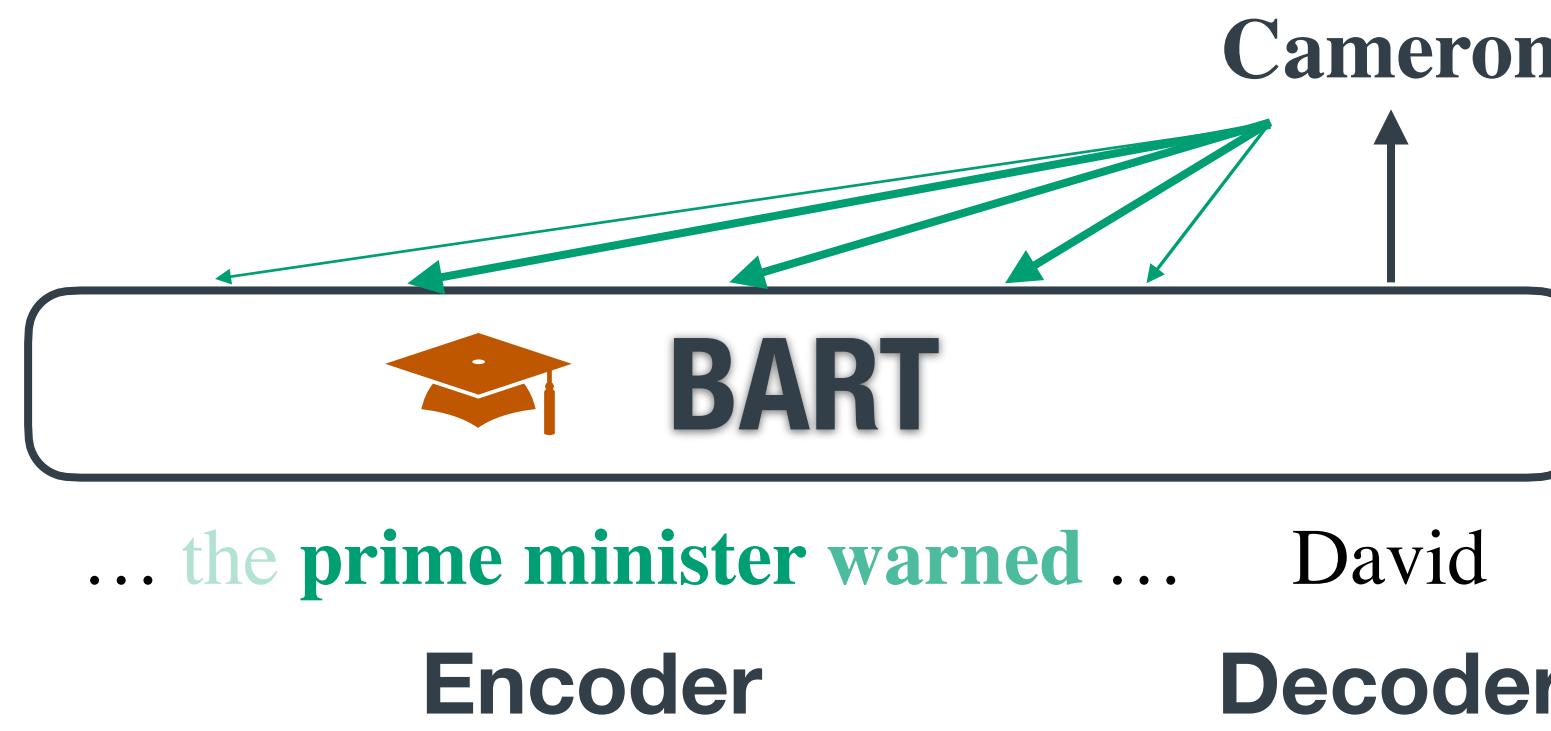
Assign score for each token according to attribution

Rank	Token	Score
1	prime	0.36
2	minister	0.21
3	warned	0.17
...

Sort them according to the score and select top-k



Evaluation Protocol



Assign score for each token according to attribution

Rank	Token	Score
1	prime	0.36
2	minister	0.21
3	warned	0.17
...

Sort them according to the score and select top-k

Input	Prefix	$p(\text{Cameron})$
prime		0.97
prime minister	+ + David	0.96
he talked to		0.08

Does the selected set of tokens help recover the prediction?



Conclusion

Why do we do ablation before attribution?

- It identifies generation modes and allow us to deploy different tools on each mode.

Can you extend the framework to other NLG tasks?

- Yes!

How effective and accurate are attribution methods?

- Fine for many cases, but still a long way to go.