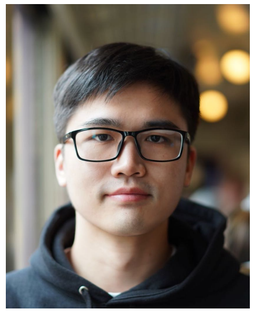
PRINCETON NLP

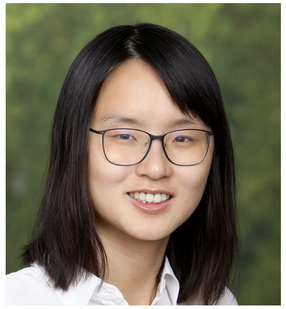
Should You Mask 15% in Masked Language Modeling?

Alexander Wettig*

Tianyu Gao*



Zexuan Zhong



Danqi Chen

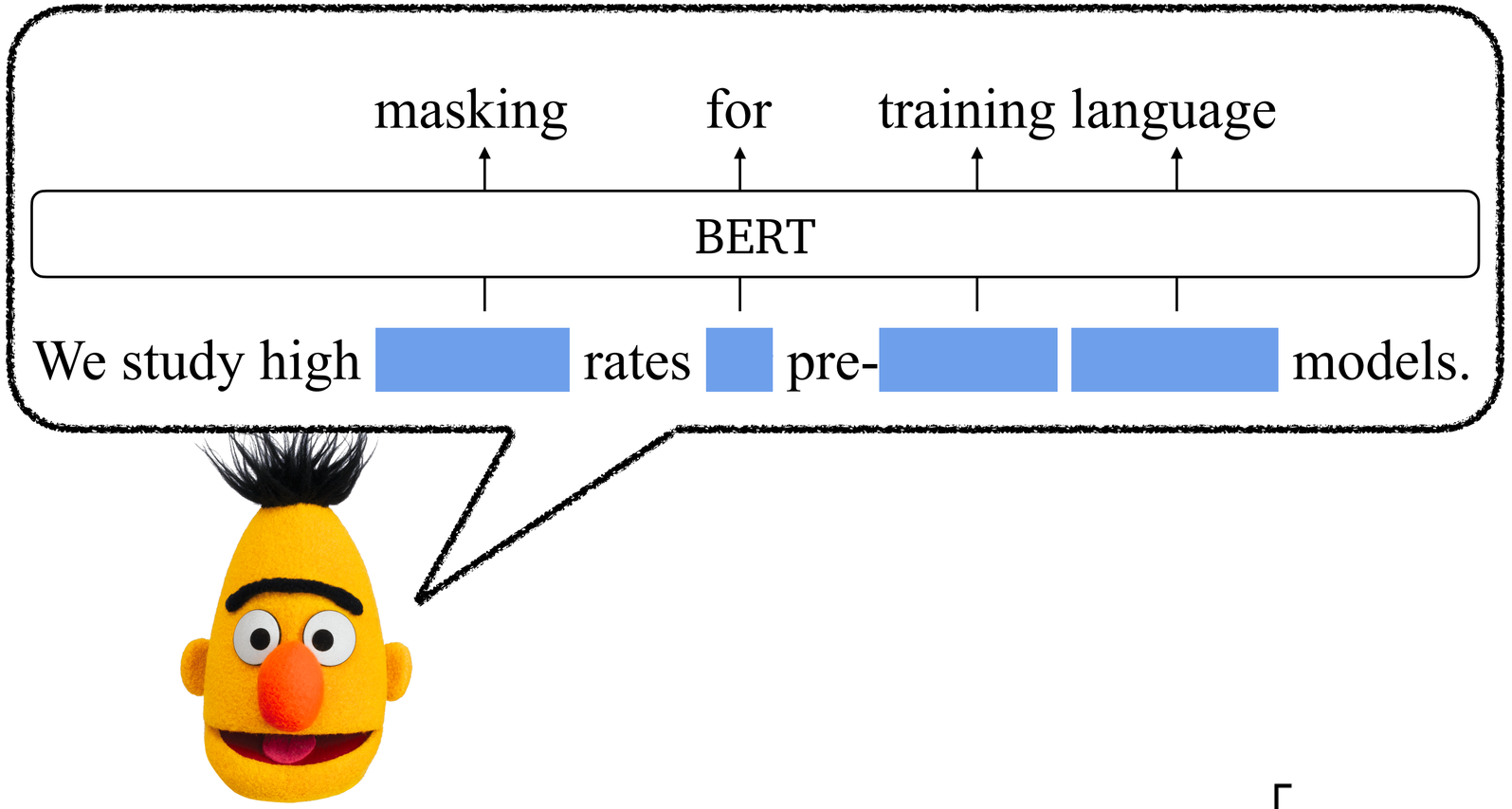
* equal contribution

BERT

We study high masking rates for pre-training language models.

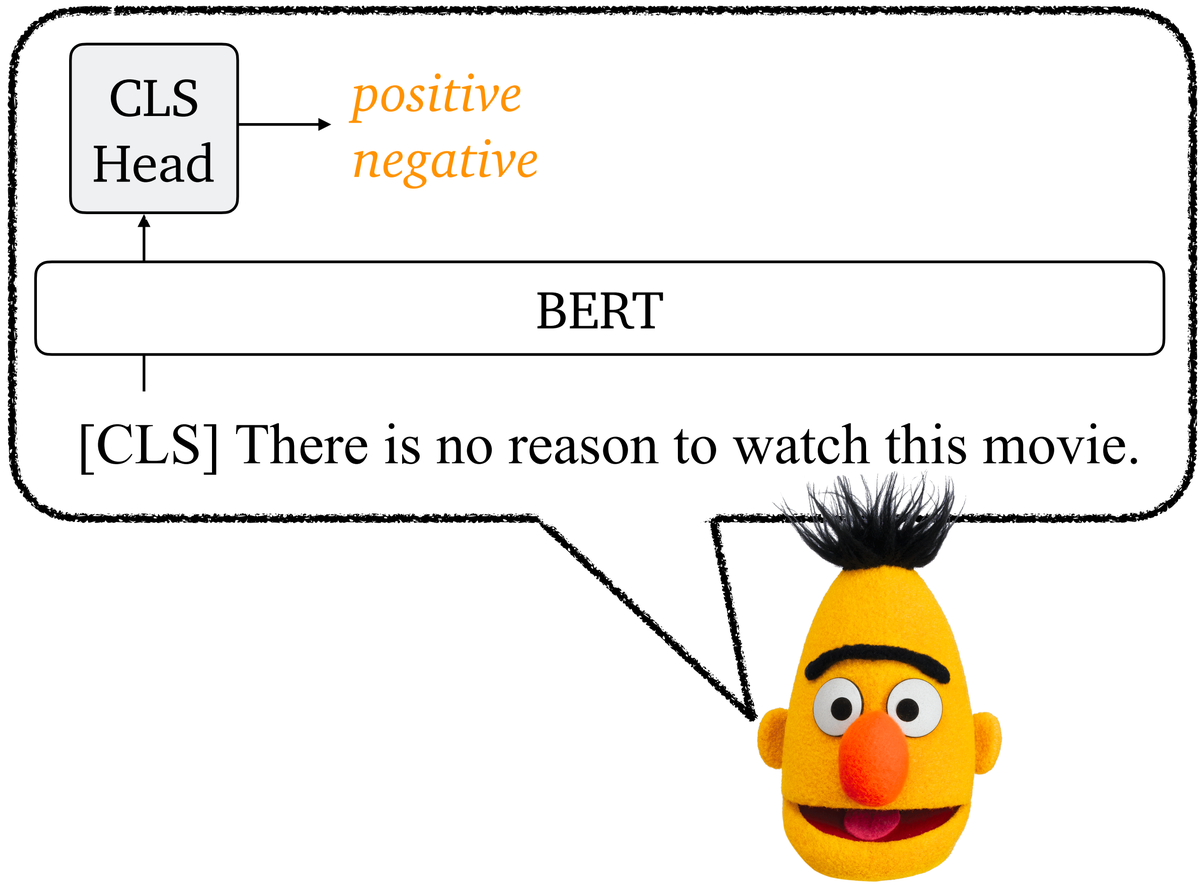
BERT

(Devlin et al., 2019)



BERT
(Devlin et al., 2019)

$$L(\mathcal{C}) = - \mathop{\mathbb{E}}_{x \in \mathcal{C}} \mathop{\mathbb{E}}_{\mathcal{M} \subset x} \left[\sum_{x_i \in \mathcal{M}} \log p(x_i | \tilde{x}) \right]$$



BERT

(Devlin et al., 2019)

RoBERTa
(Liu et al., 2019)

SpanBERT
(Joshi et al., 2019)

DeBERTa
(He et al., 2020)

BERT
(Devlin et al., 2019)

PMI-Masking
(Levine et al., 2021)

ALBERT
(Lan et al., 2020)

RoBERTa
(Liu et al., 2019)

SpanBERT
(Joshi et al., 2019)

DeBERTa
(He et al., 2020)

BERT
(Devlin et al., 2019)

mask 15% of tokens

PMI-Masking
(Levine et al., 2021)

ALBERT
(Lan et al., 2020)

RoBERTa
(Liu et al., 2019)

SpanBERT
(Joshi et al., 2019)

DeBERTa
(He et al., 2020)

BERT
(Devlin et al., 2019)

mask 15% of tokens

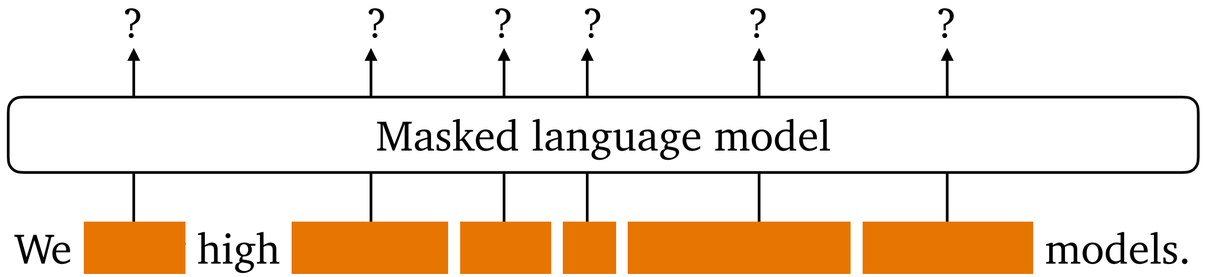
PMI-Masking
(Levine et al., 2021)

ALBERT
(Lan et al., 2020)

Why 15% masking?

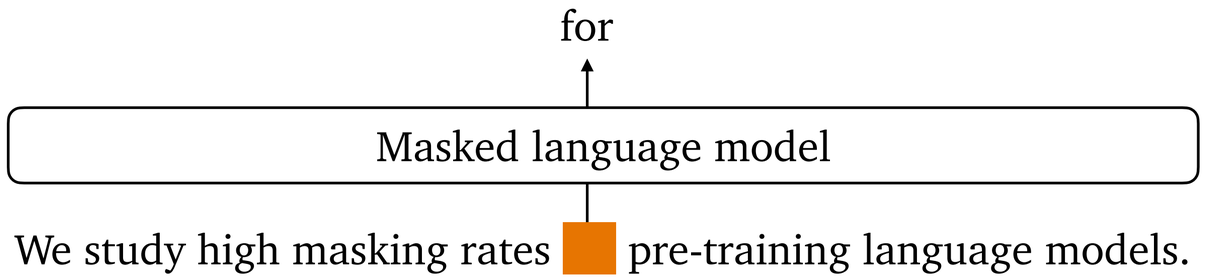
The common belief is

- Masking **more** than 15%:



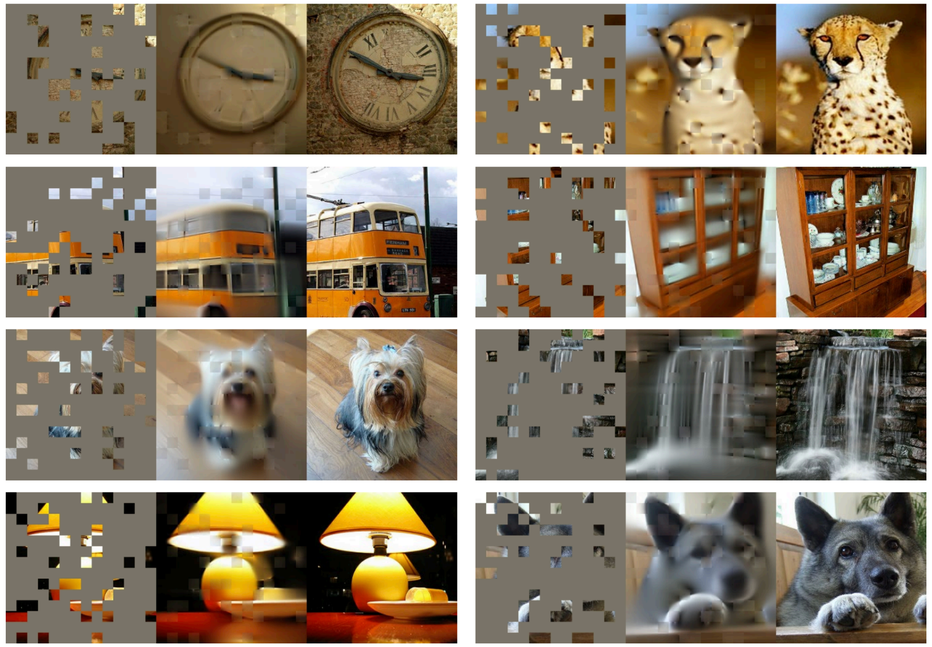
➡ Insufficient context to learn, reconstruction becomes impossible

- Masking **less** than 15%:



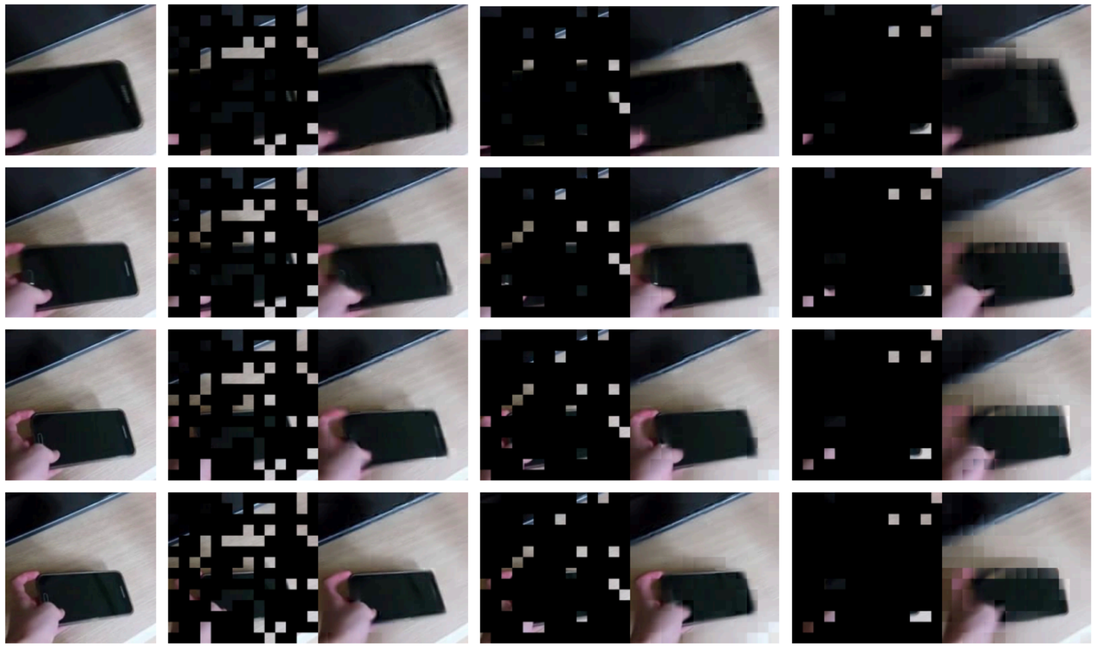
➡ Model learns from fewer predictions, training is inefficient

Masking in images and videos



MAE (He et al., 2021)

Mask 75%

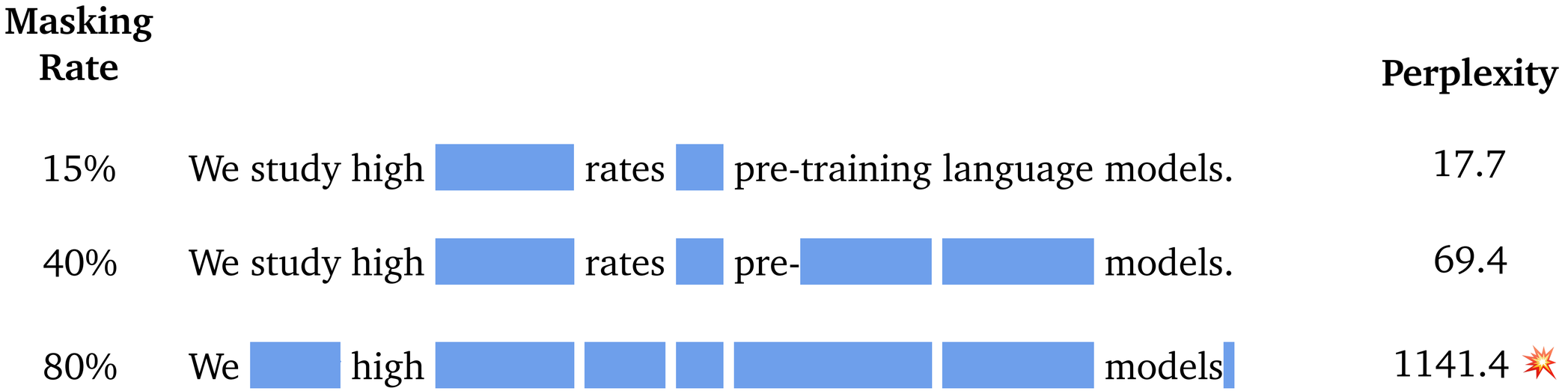


VideoMAE (Tong et al., 2022)

Mask 90-95%

Different **information density** between language and vision

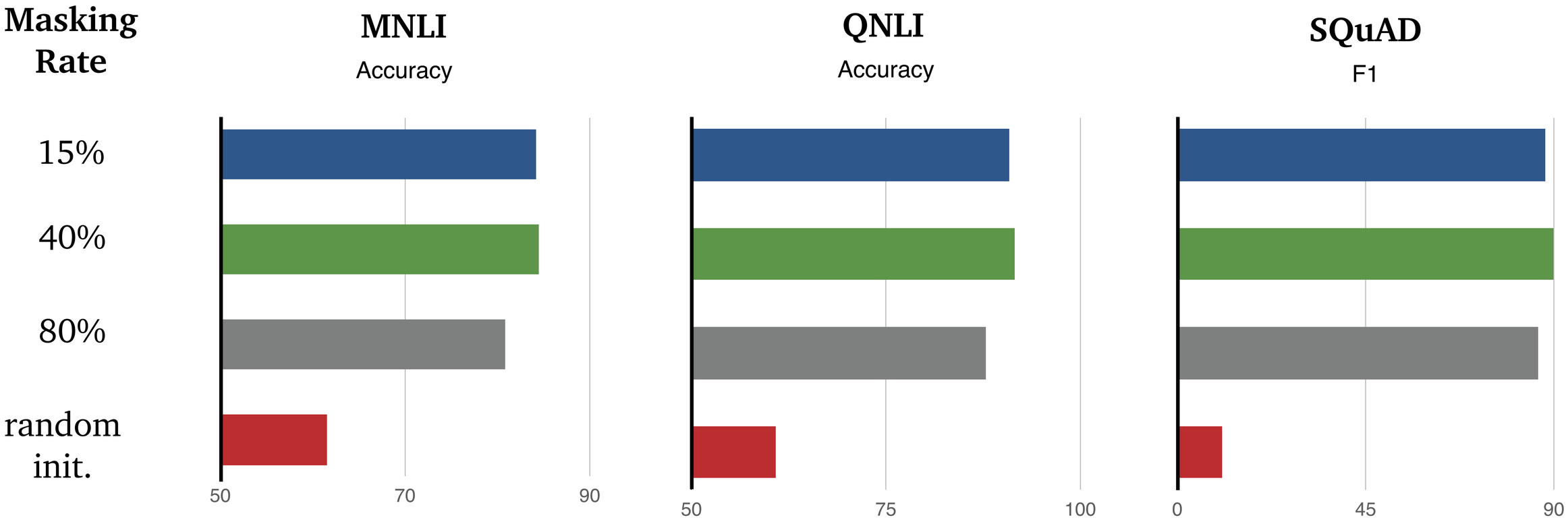
Varying the Masking Rate



All results reported, unless specified, are large models trained with the efficient pre-training recipe (Izsak et al., 2021).
Perplexities are evaluated on the validation set of one billion word benchmark (Chelba et al., 2013) at the corresponding masking rate of each model.

Varying the Masking Rate

Fine-tuning on Downstream Task Evaluation

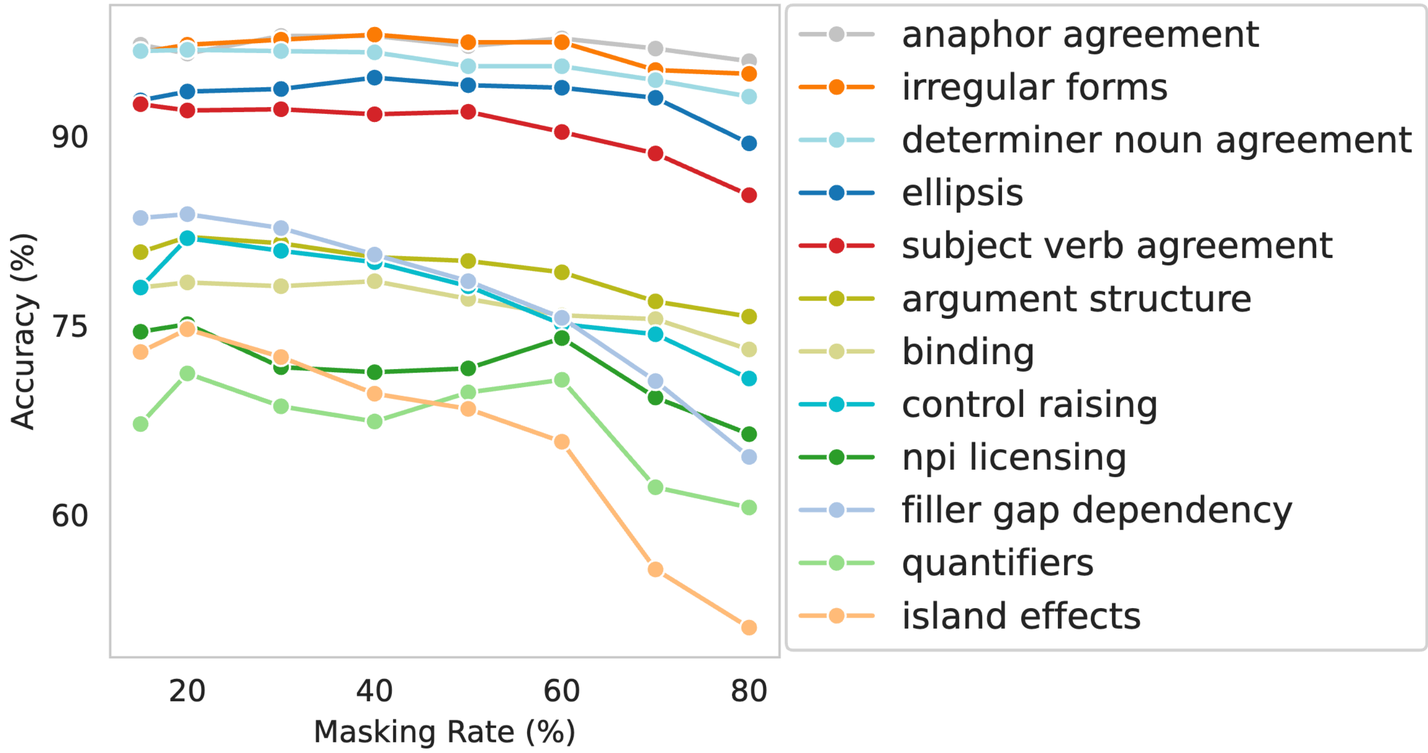


- Hypothesis: Even if reconstructing tokens is impossible due to high masking rates, BERT can still learn good initialization for downstream tasks

For SQuAD, we continue training the models with 512-token sequences for 2,300 steps.

Varying the Masking Rate

Linguistic Probing on BLIMP



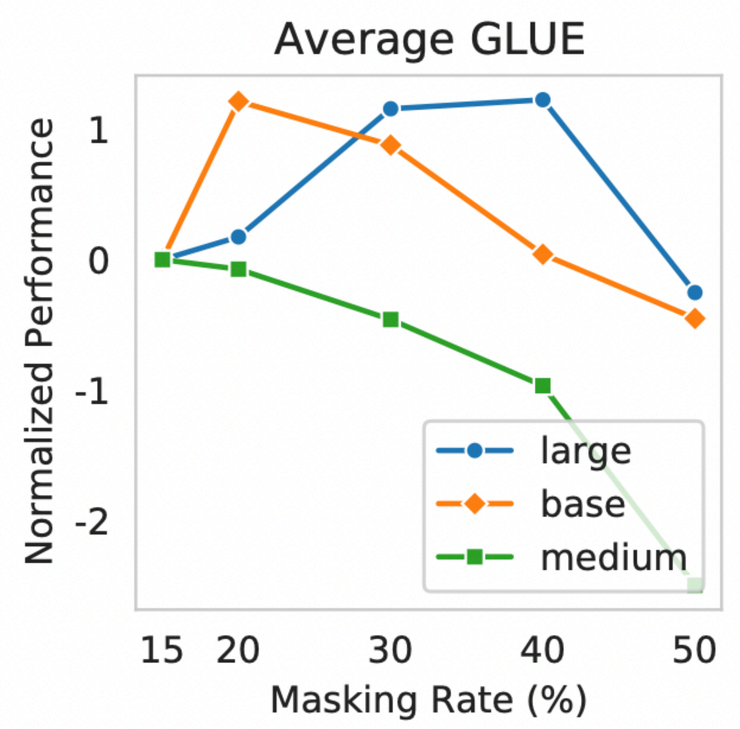
Even 80% learns non-trivial linguistic features!

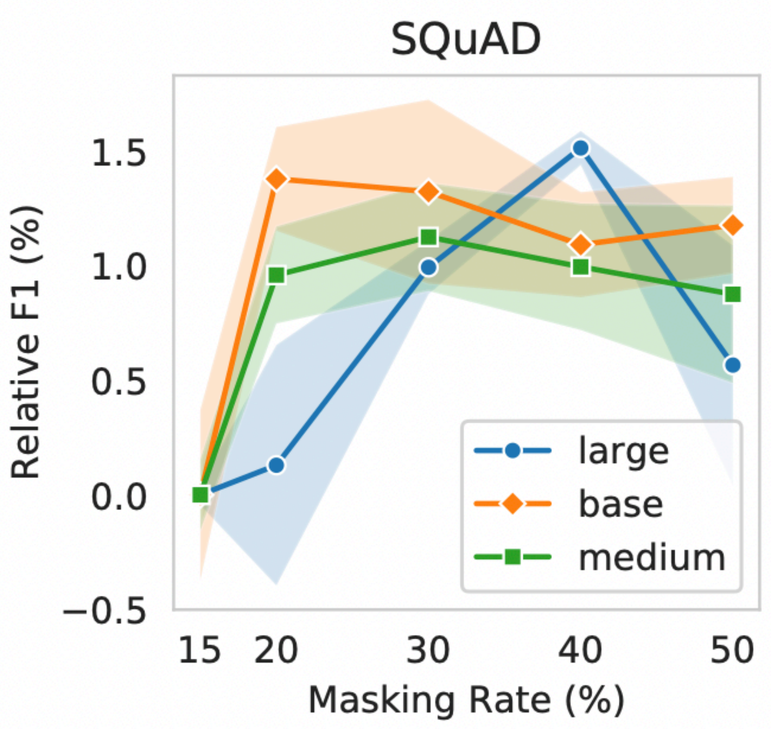
We evaluate using pseudo log-likelihood scoring (Salazar et al., 2020).

What factors influence the optimal masking rate?

Masking Rate vs. Model Size

We pre-train and fine-tune `medium` (51M) > `base` (124M) > `large` (354M) models





- Hypothesis: larger models can handle the **harder pre-training task** better

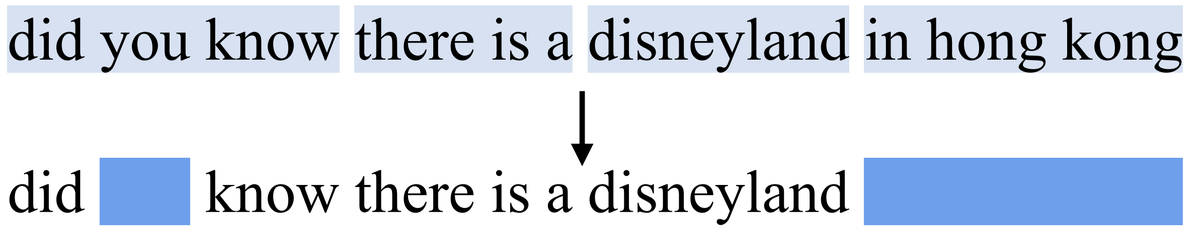
Masking Rate vs. Masking Strategy

did you know there is a disneyland in hong kong

Uniform (Joshi et al., 2020; Raffel et al., 2020)

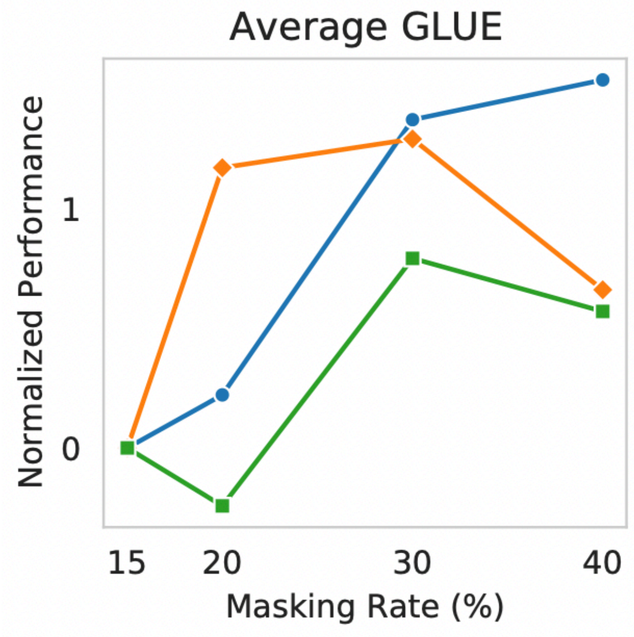
Span Masking (Joshi et al., 2020; Raffel et al., 2020)

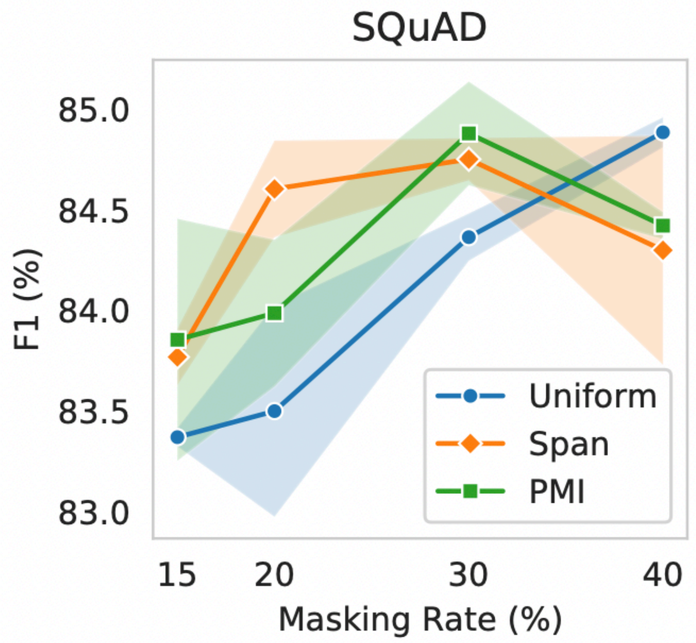
PMI masking (Levine et al., 2021)



Masking Rate vs. Masking Strategy

- Span Masking (Joshi et al., 2020; Raffel et al., 2020) or PMI masking (Levine et al., 2021) create more challenging masking patterns to avoid local cues such as “Hong [MASK]”
- These strategies outperform uniform masking at a masking rate of 15%
- However, **uniform masking** has a **higher** optimal masking rate and can be **competitive**.



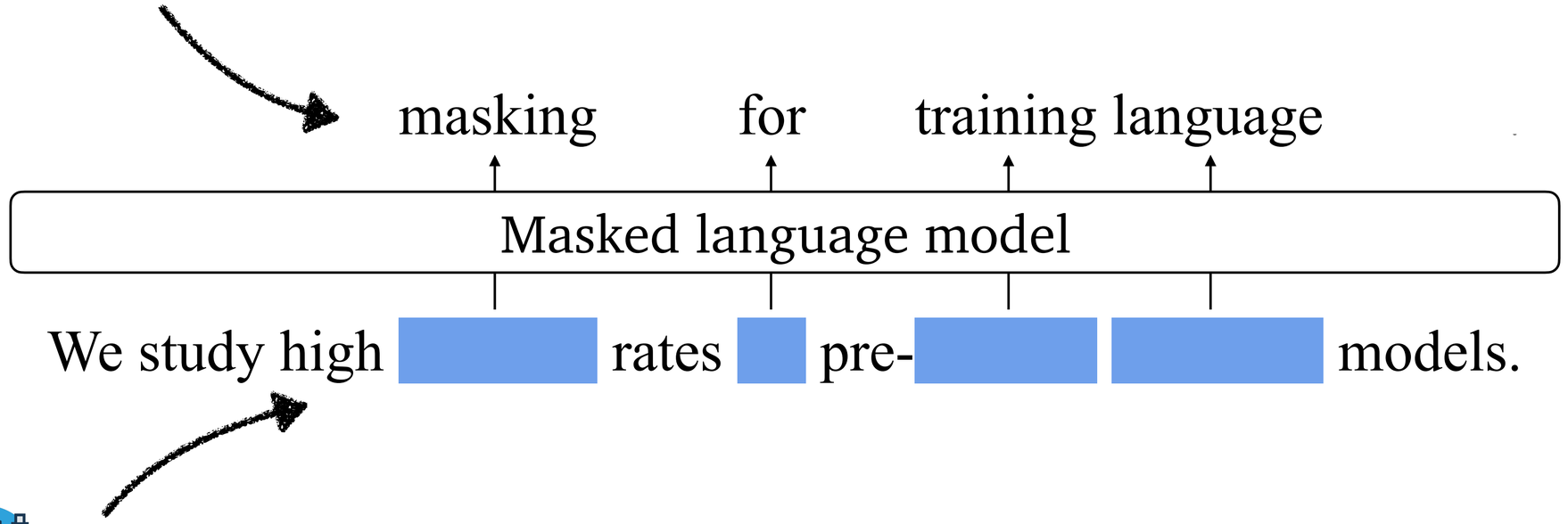


How do models benefit from high masking rates?

Corruption vs. Prediction

Prediction rate affects optimization.

Higher prediction rate leads to more training signals and benefits training efficiency 👍



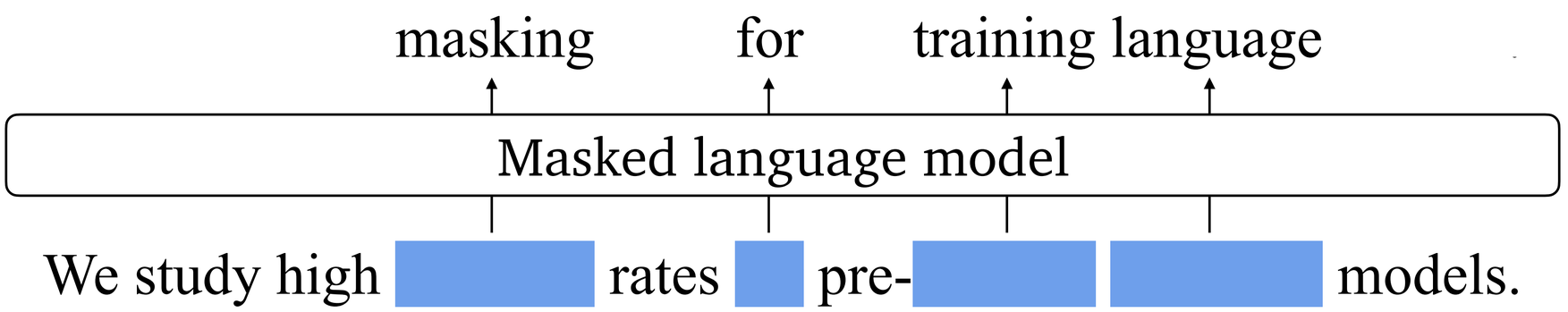
Corruption rate controls task difficulty.

Higher corruption rate leads to harder task, which models may not be able to learn from 👎

Corruption vs. Prediction

Prediction rate affects optimization.

Higher prediction rate leads to more training signals and benefits training efficiency 👍

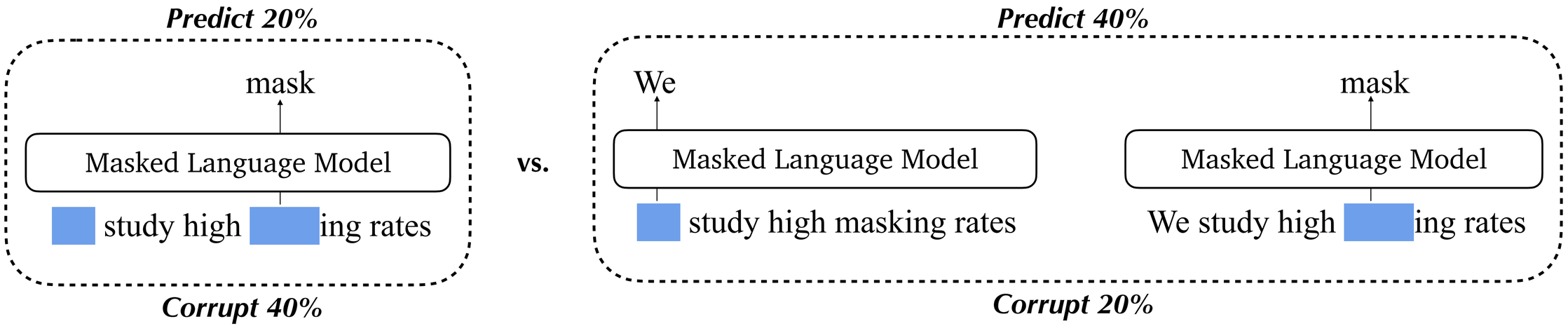


Corruption rate controls task difficulty.

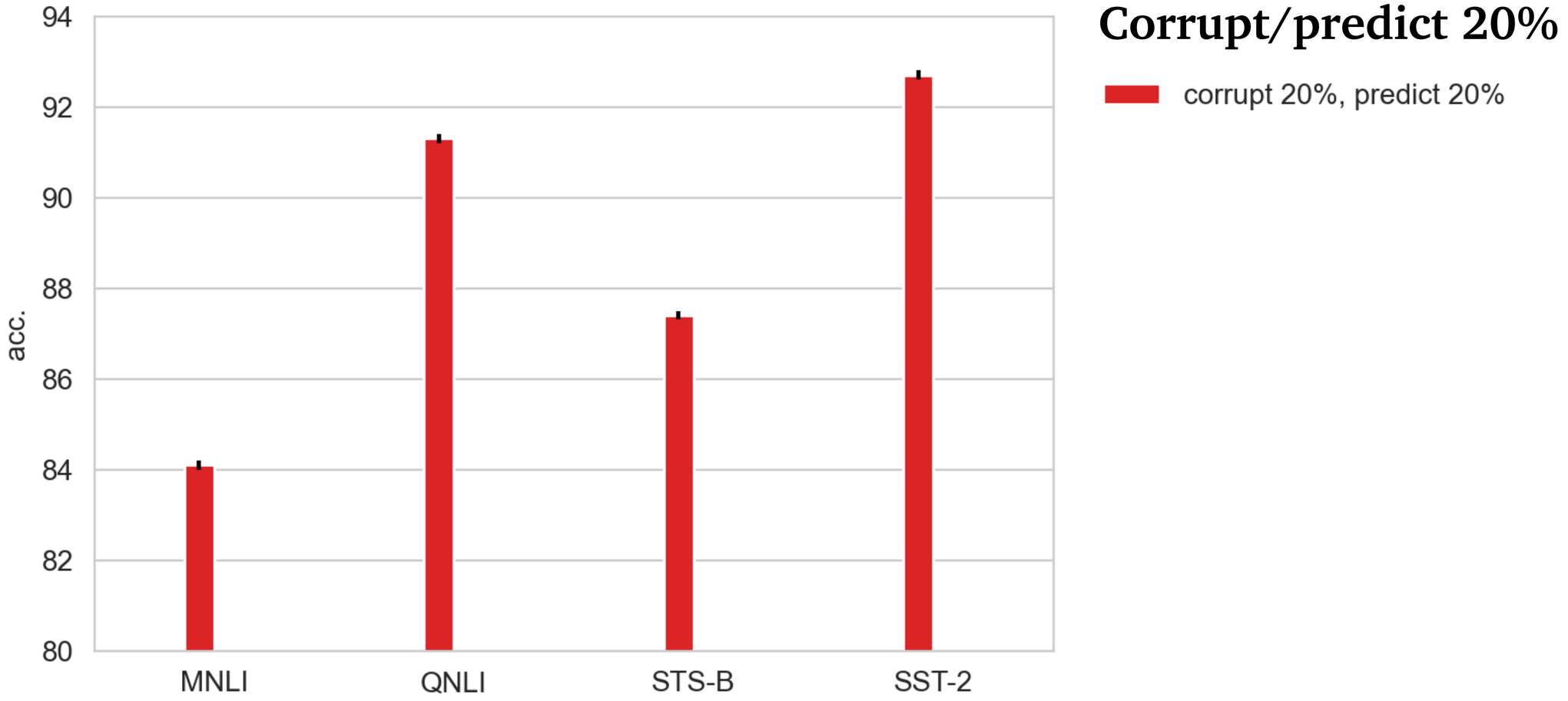
Higher corruption rate leads to harder task, which models may not be able to learn from 👎

Disentangling Corruption and Prediction

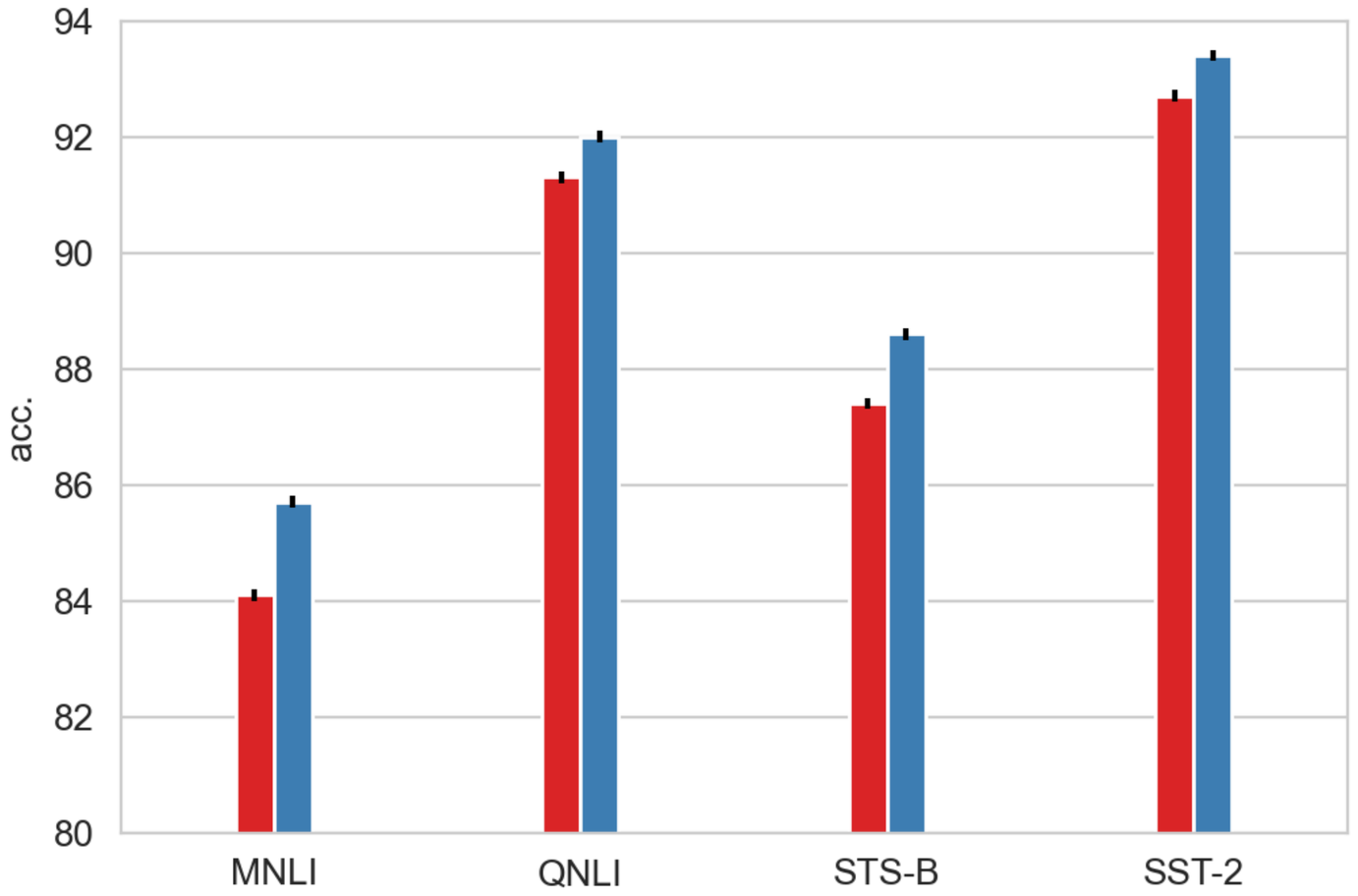
- We design experiments to disentangle **corruption** and **prediction**.



Corruption vs. Prediction



Corruption vs. Prediction

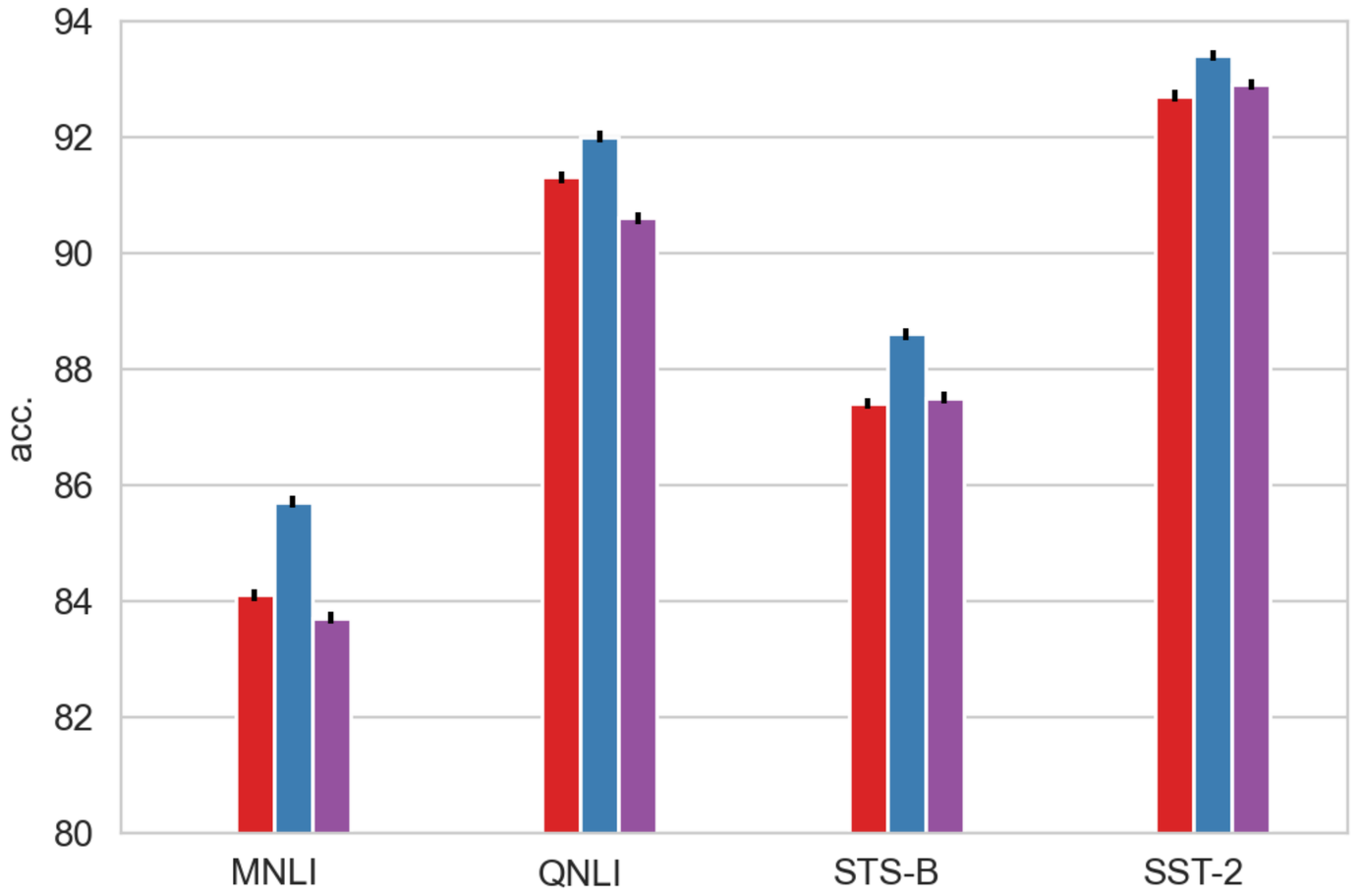


Corrupt 20% → Predict 40%

corrupt 20%, predict 20%

corrupt 20%, predict 40%

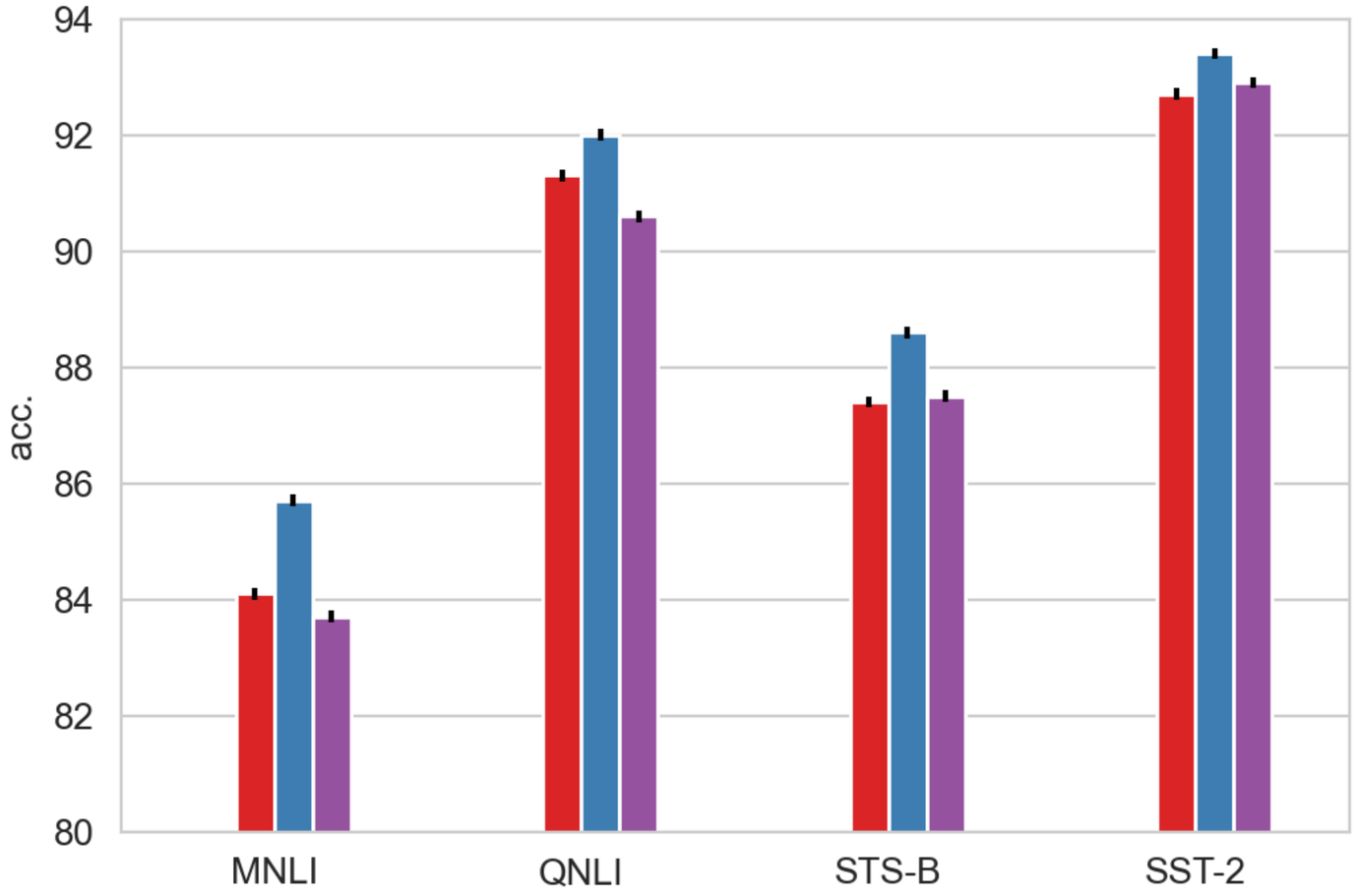
Corruption vs. Prediction



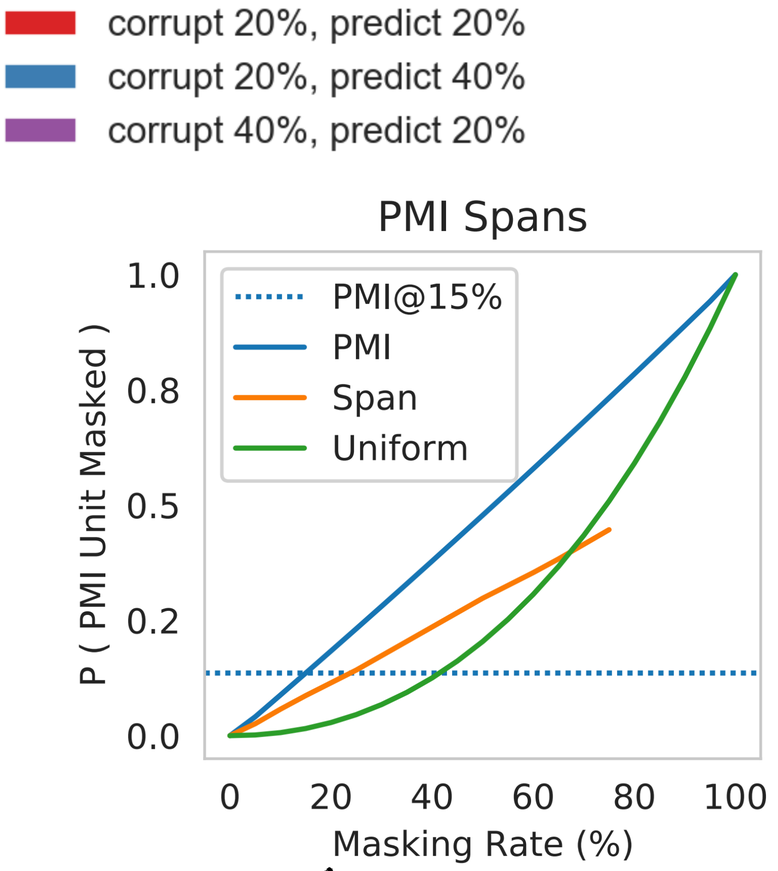
Corrupt 40% → Predict 20%

corrupt 20%, predict 20%
corrupt 20%, predict 40%
corrupt 40%, predict 20%

Corruption vs. Prediction

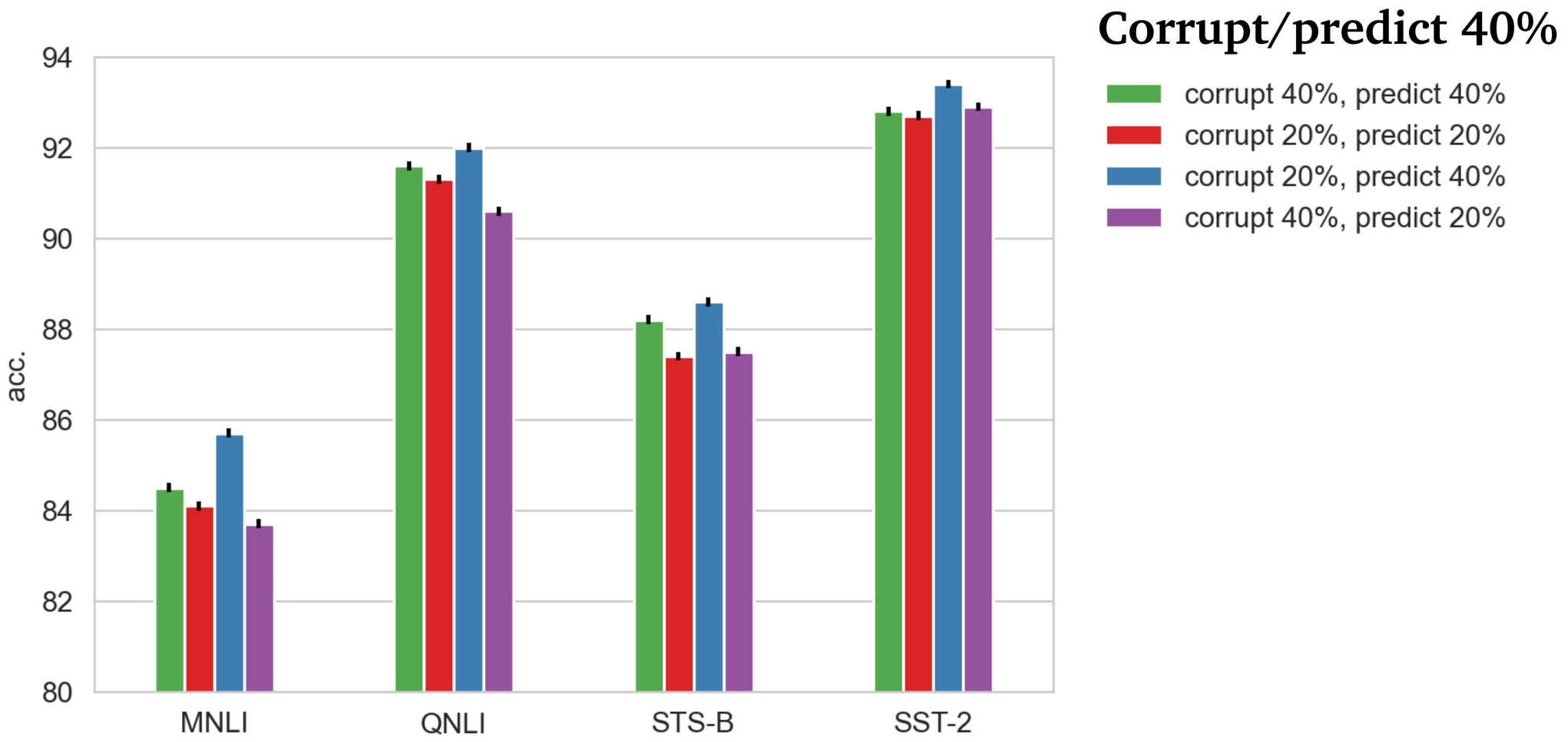


Corrupt 40% → Predict 20%



No observable benefits from regularization at higher masking rates

Corruption vs. Prediction



Benefit from higher **prediction rate** > Penalty from higher **corruption rate**

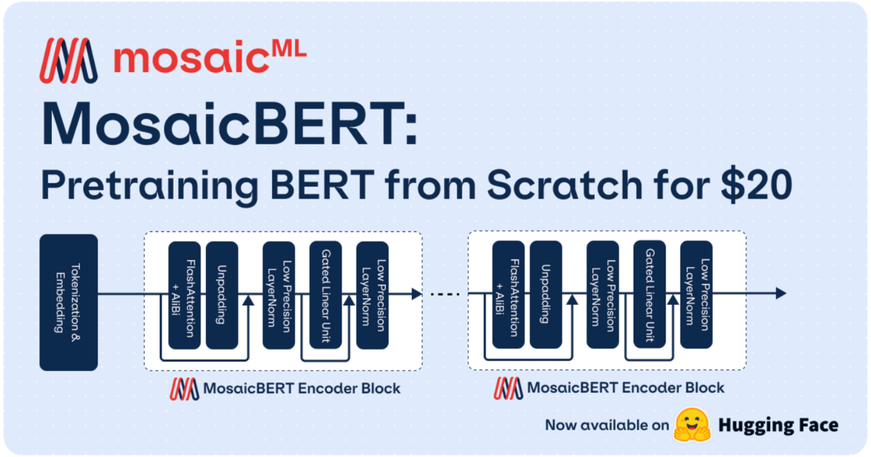
Corruption Strategy: 80-10-10 Rule Is Not Necessary

- 80-10-10 rule from BERT (Devlin et al., 2019)
  - 80% of masked tokens with [MASK]
  - 10% with a random token
  - 10% with the original token

	MNLI	QNLI	QQP	STS-B	SST-2
40% mask	$84.5_{0.1}$	$91.6_{0.1}$	$88.1_{0.0}$	$88.2_{0.1}$	$92.8_{0.1}$
+5% same	84.2↓	91.0↓	87.8↓	88.0↓	93.3↑
w/ 5% rand	84.5	91.3↓	87.9↓	87.7↓	92.6↓
w/ 80-10-10	84.3↓	91.2↓	87.9↓	87.8↓	93.0↑

Masking all performs the best!

Implications



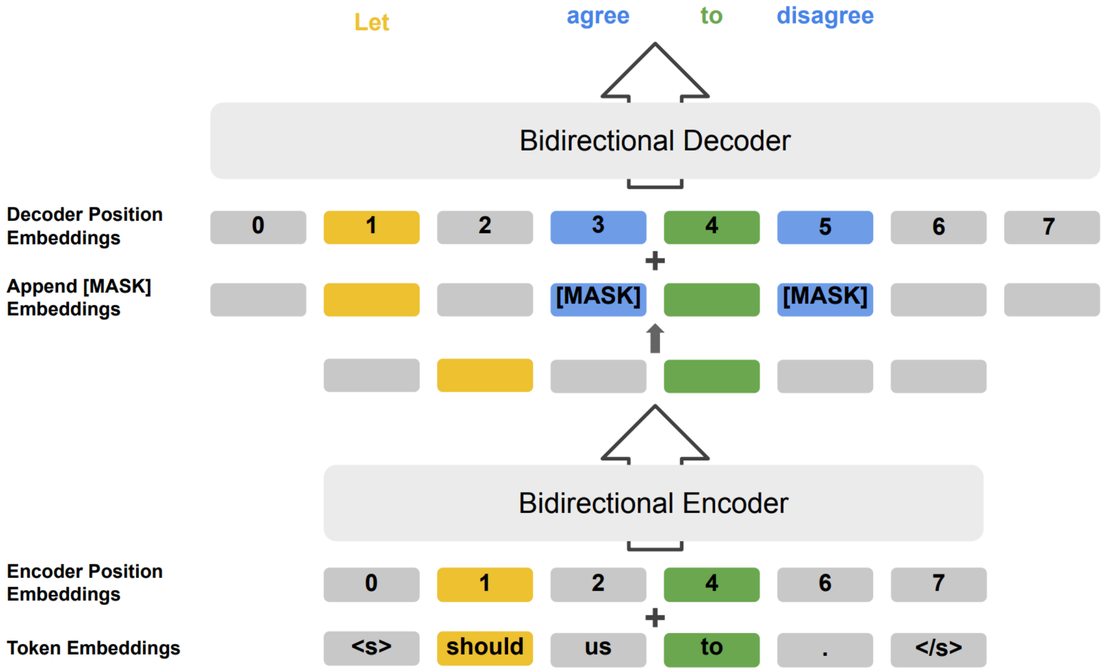
30% masking ratio adapted by industry for efficient pre-training

Can encode masks only in a decoder step to achieve efficiency gains.

Liao et al., 2022, also Meng et al., 2023

"Mask More and Mask Later: Efficient Pre-training of Masked Language Models by Disentangling the [MASK] Token"

More masking ➡ Compute savings



Thank you!

Contact:

awettig@cs.princeton.edu

tianyug@cs.princeton.edu

@princeton_nlp