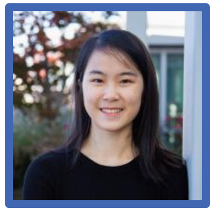
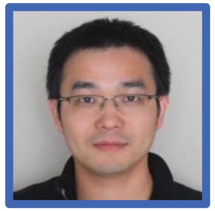
On the Influence of Masking Policies in Intermediate Pre-training


Qinyuan Ye[1,†]
† Work partially done while at Facebook AI


Belinda Z. Li[2,†]
† Work partially done while at Facebook AI


Sinong Wang[3]


Benjamin Bolte[3]


Hao Ma[3]


Wen-tau Yih[3]


Xiang Ren[1]


Madian Khabsa[3]

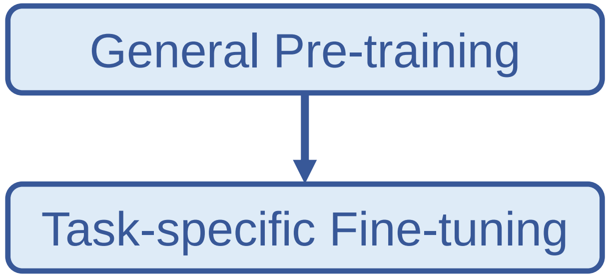
1 USC Viterbi Department of Computer Science

2 MIT CSAIL

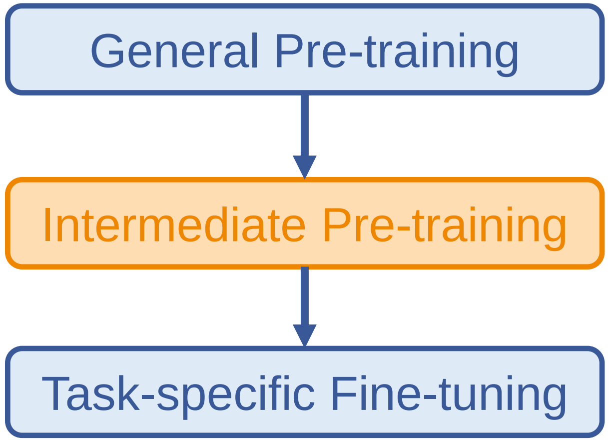
3 facebook Artificial Intelligence

Motivation

- ***"Pre-train then fine-tune"*** is a predominant pipeline in NLP.



- Inserting an ***intermediate stage with a hand-crafted masking policy*** can be helpful.

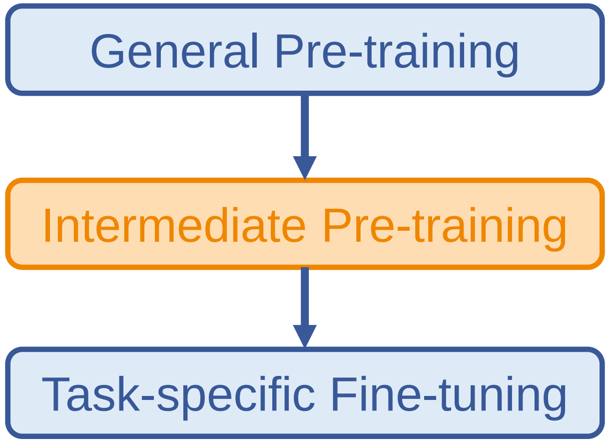


masking named entities → better closed-book QA models (Roberts et al., 2020)

Motivation

- In this work we offer a large-scale ***empirical study*** to investigate ***the influence of masking policies in intermediate pre-training***.

- We aim to answer
  - ***In what cases*** such intermediate pre-training is helpful
  - Whether hand-crafted heuristic objectives are ***near-optimal***
  - Whether a masking policy designed for one task is ***generalizable*** beyond that task



Preliminaries – Masked Language Modeling

- Masked language modeling (MLM) and variants are common for pre-training large-scale transformers, e.g., BERT, RoBERTa, BART, T5.

Original Sentence

The quick brown fox jumped over the lazy dog.

Masking Policy

Masked Sentence

The quick **<mask>** fox jumped **<mask>** the lazy dog.

Text-to-text Model

Pre-train Target

The quick **brown** fox jumped **over** the lazy dog.

Analysis Setup – Training Pipeline

General Pre-training with Random Masks

Source: In Newtonian physics, free fall <mask> motion of a <mask> where gravity is <mask> only <mask> upon it.

Text-to-text Model

Target: In Newtonian physics, free fall is any motion of a body where gravity is the only force acting upon it.

Intermediate Pre-training

Source: In Newtonian physics, free fall is any motion of a body where <mask> is the only force acting upon it.

Text-to-text Model

Target: gravity

Task-specific Fine-tuning

Source: which force acts on an object in free fall

Text-to-text Model

Target: gravity

We ensure that the masking policy is the only variable in this pipeline.

Analysis Setup – Downstream Tasks

Closed-book QA

TQA: TriviaQA
WQ: WebQuestions
NQ: Natural Questions

Knowledge-intensive Tasks

AY2: Entity Linking
ZSRE: Zero-shot Relation Extraction
WoW: Dialogue

Multiple-choice QA

WIQA
QuaRTz
ROPES

Each example is a source-target pair *(s, t)*, accompanied with a context paragraph *c*

Compared Masking Policies (1/3)

- **Heuristics**

Original Objective (+Orig)
Mask 15% randomly selected tokens and recover full sequence

Masked Sentence
I **<mask>** lived in Los **<mask>** for two years.

Text-to-text Model

Pre-train Target
I have lived in Los Angeles for two years.

Random Masking (+Rand)
Mask 15% randomly selected tokens and recover the masked tokens

Masked Sentence
I **<mask>** lived in Los **<mask>** for two years.

Text-to-text Model

Pre-train Target
lived Angeles

Salient Span Masking (+SSM)
Mask and recover one named entity

Masked Sentence
I have lived in <mask> for two years.

Text-to-text Model

Pre-train Target
Los Angeles

Compared Masking Policies (2/3)

- **Supervised**
  - Identify ***likely-answers*** and mask them
  - The masking policies is similar to an extractive reading comprehension model
  - The policy is trained with ***(context, answer)*** examples, ***without the question***

Example

Input: [Charles, Schulz, was, the, creator, of, Snoopy]

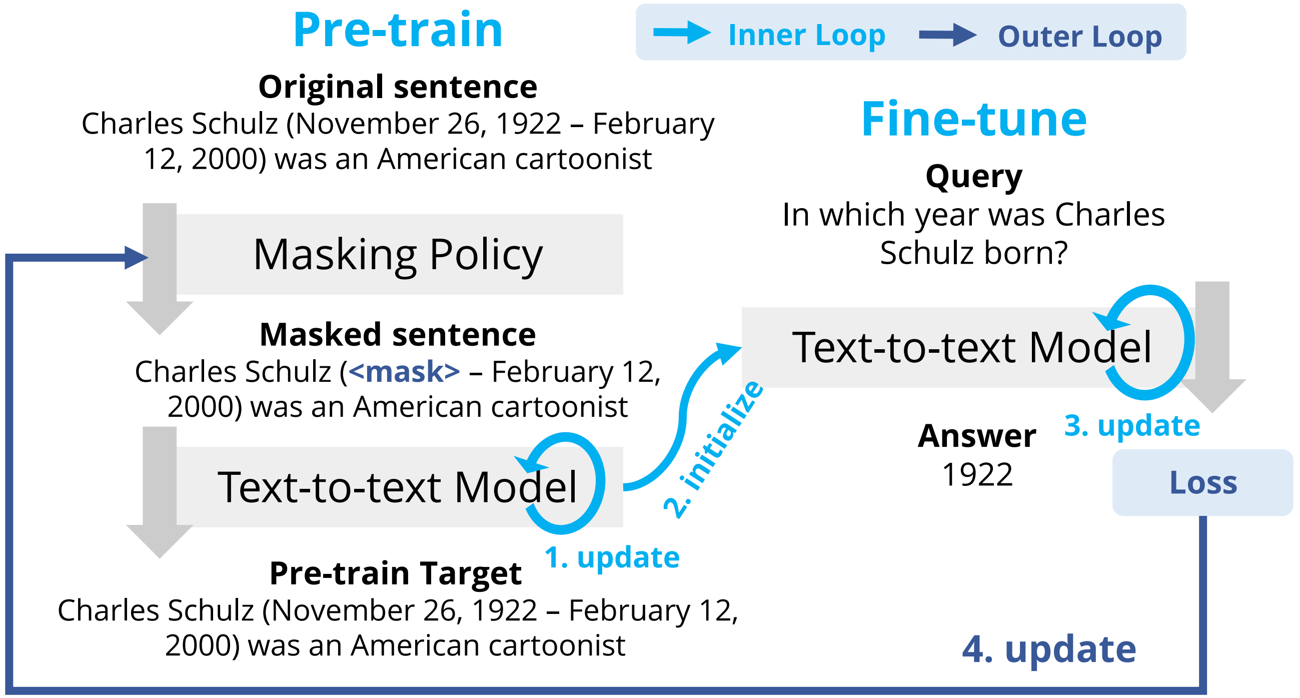
Masking Policy

Start Index: [1,0,0,0,0,0,0]
End Index: [0,1,0,0,0,0,0]

Compared Masking Policies (3/3)

Heuristic, Supervised, __Meta-learned__

- **Meta-learned**
  - Our goal is closely related to the concept of "***learning to learn***" (Schmidhuber, 1987; Thrun and Pratt, 1998).
  - The masking policy should help the text-to-text to learn quickly when fine-tuned on downstream tasks.



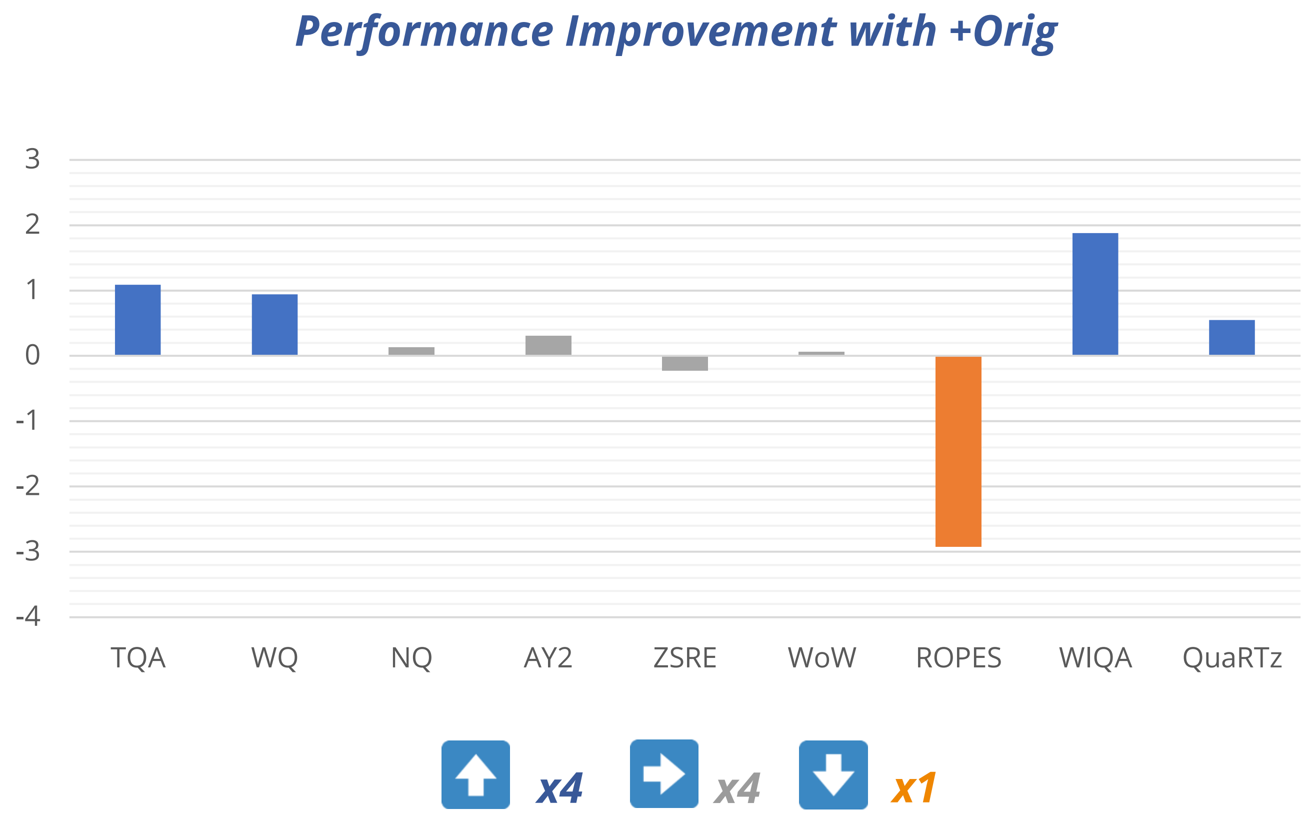
A more detailed algorithm box can be found in the paper!

Algorithm 1 Meta-learning Policy $g(.;\phi)$

Input: Dataset $\mathcal{S} = \{(\mathbf{c},\mathbf{s},\mathbf{t})\}$
Output: Masking Policy $g(.;\phi)$, A Pre-trained Language Model $f(.;\theta)$

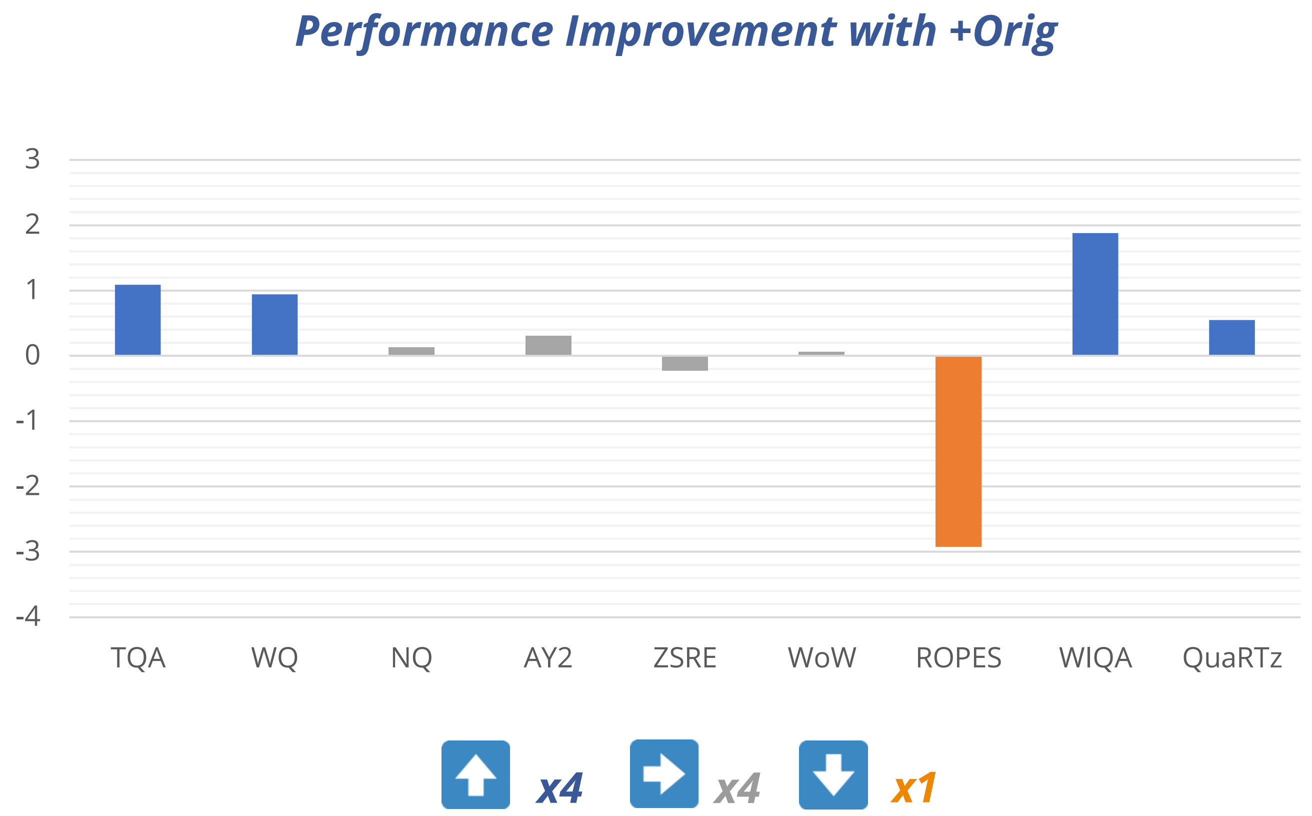
for $p = 1..T$ **do**
$\{(\mathbf{c},\mathbf{s},\mathbf{t})\} = \text{SampleBatch}(\mathcal{S})$
$\mathbf{xc};\mathbf{d} = g(\mathbf{x};\phi^{(p)}); \mathbf{x}' = \mathbf{x} \oplus \mathbf{d}$ // Apply current masking policy $g(.;\phi)$
$\theta^{(p,1)} = \theta^{(p,0)} - \alpha_0 \nabla_{\theta^{(p,0)}} \mathcal{L}(f(\mathbf{x}';\theta^{(p,0)}),\mathbf{x})$ // Inner Loop Update 1: Pre-train on $(\mathbf{x},\mathbf{x}')$
$\theta^{(p,2)} = \theta^{(p,1)} - \alpha_1 \nabla_{\theta^{(p,1)}} \mathcal{L}(f(\mathbf{s};\theta^{(p,1)}),\mathbf{t})$ // Inner Loop Update 2: Fine-tune on $(\mathbf{s},\mathbf{t})$
$\mathcal{L}' = \mathcal{L}(f(\mathbf{s};\theta^{(p,2)}),\mathbf{t}) - \mathcal{L}(f(\mathbf{s};\theta^{(p,1)}),\mathbf{t})$ // Compute loss measuring how fast the model adapts to $(\mathbf{s},\mathbf{t})$
$\mathcal{L}_{reg} = \delta[|\gamma l(\mathbf{x}) - l(\mathbf{d})| > \epsilon l(\mathbf{x})](\gamma l(\mathbf{x}) - l(\mathbf{d}))^2$ // Compute regularization loss to control masking budget
$\phi^{(p+1)} = \phi^{(p)} - \alpha_2 \nabla_{\phi^{(p)}}(\mathcal{L}' + \beta\mathcal{L}_{reg}(\mathbf{x},\mathbf{d}))$ // Outer Loop Update for ϕ
$\theta^{(p+1,0)} = \theta^{(p,1)}$ // Maintain pre-train progress at timestamp p

Analysis – Comparing Heuristic Policies



Overall, intermediate pre-training with the original objective leads to improved performance.

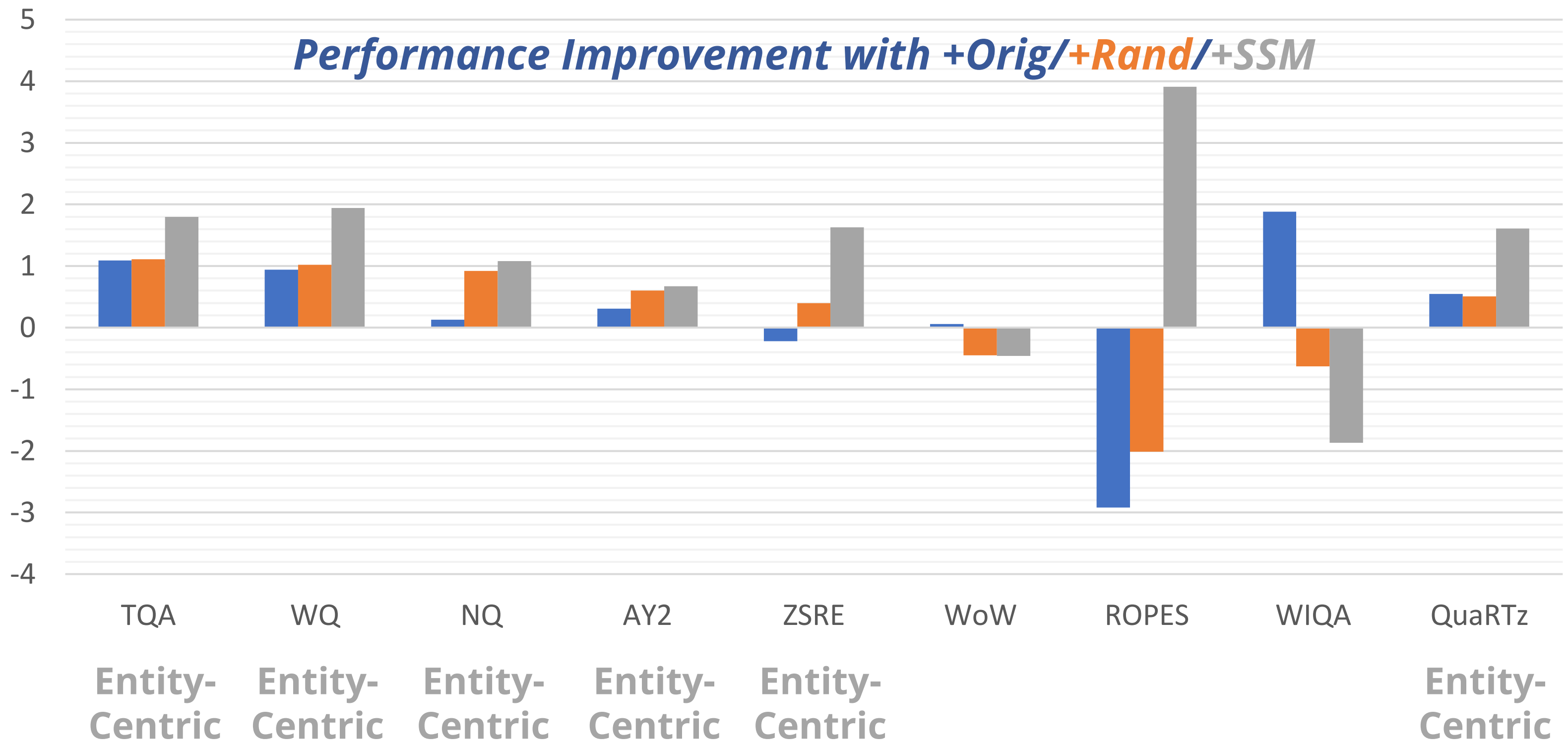
Analysis – Comparing Heuristic Policies



ROPES dataset is ***the only exception***. We found that intermediate pre-training may lead to ***catastrophic forgetting***.

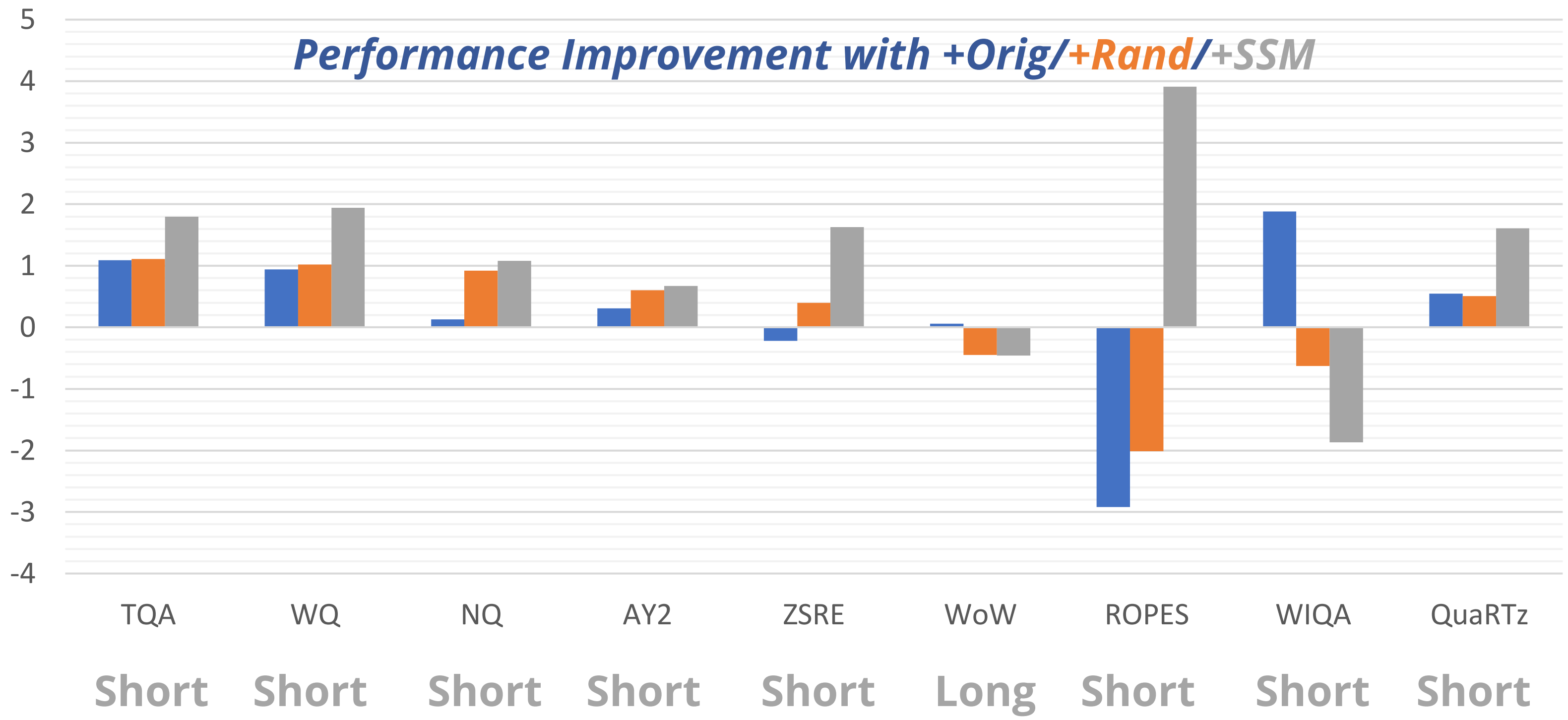
Pay attention to the corpus from which the dataset is created!

Analysis – Comparing Heuristic Policies



- SSM is beneficial for all entity-centric tasks.
- Use heuristic masking policies that resemble the downstream tasks, or masking information known to be important for the downstream task, tend to be helpful.

Analysis – Comparing Heuristic Policies



+Orig is better for long outputs

Masked Sentence
I **<mask>** lived in Los **<mask>** for two years.

Text-to-text Model

Pre-train Target
I have lived in Los Angeles for two years.

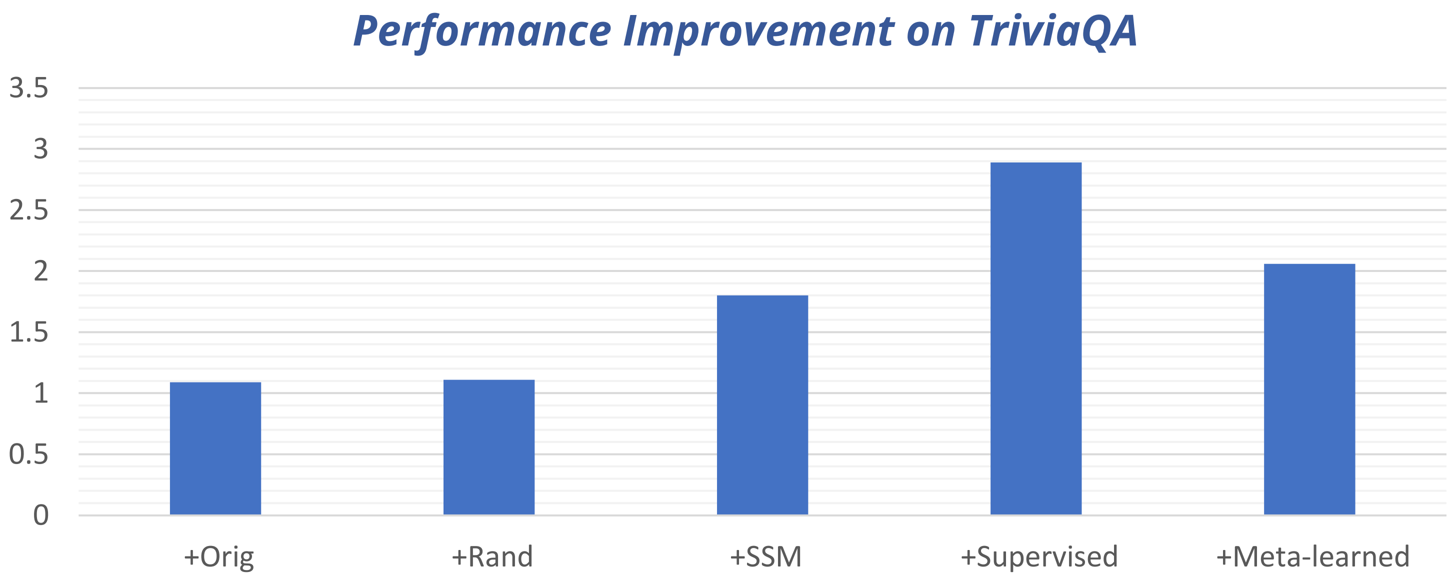
+Rand is better for short outputs

Masked Sentence
I **<mask>** lived in Los **<mask>** for two years.

Text-to-text Model

Pre-train Target
have Angeles

Analysis – Bringing in Learned Policies



Learned policies are most successful on TriviaQA.

Meta-learned policies also outperform +Orig on NQ (+0.98 EM), ZSRE (+0.32 EM) and ROPES (+10.03 Acc.)

Analysis – Bringing in Learned Policies

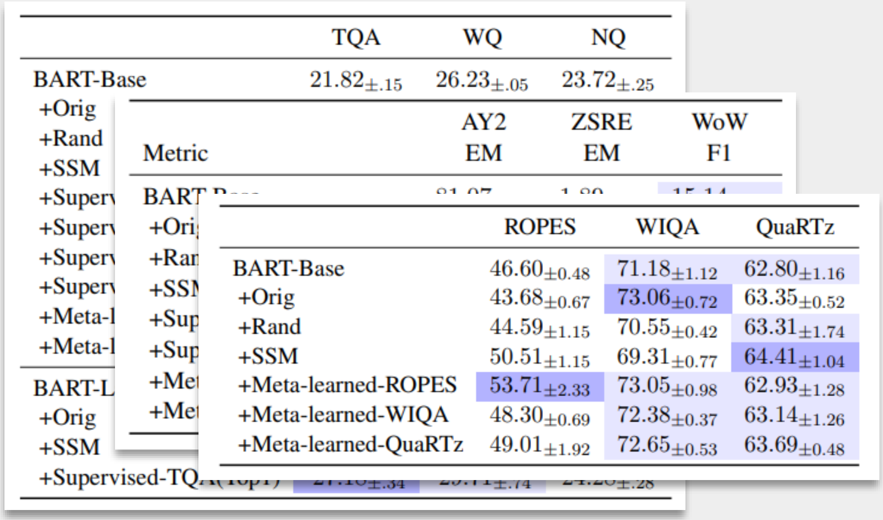
Additional Observations

Improved Learning Efficiency: Supervised policy has better learning efficiency than SSM on TriviaQA.

Generalization of Learned Policies: Learned masking policies ***can positively transfer***. However more investigation is needed.

Overfitting on ZSRE: The learned policy may ***overfit to the training data*** if there is a mismatch between train/test data

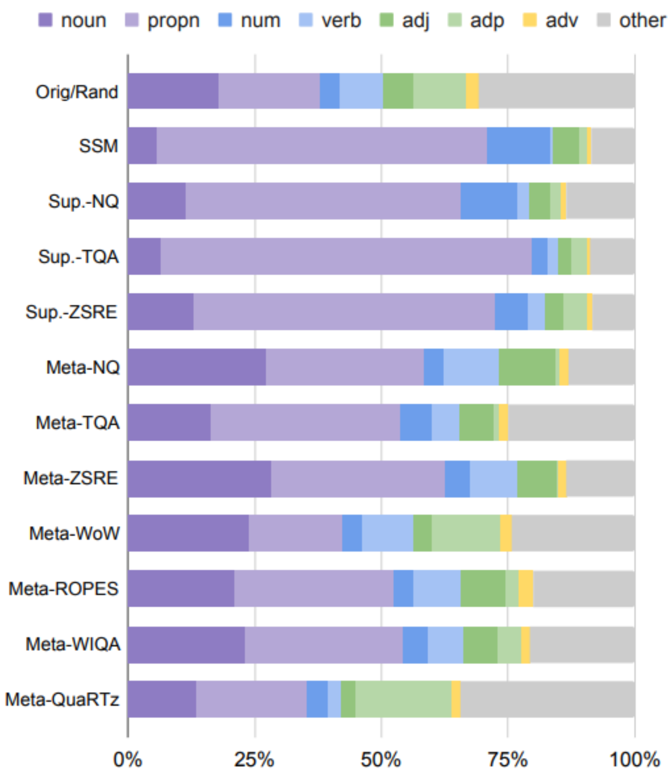
Please check out the full result tables in our paper!



	ROPES	WIQA	QuaRTz
BART-Base	$46.60_{\pm0.48}$	$71.18_{\pm1.12}$	$62.80_{\pm1.16}$
+Orig	$43.68_{\pm0.67}$	$73.06_{\pm0.72}$	$63.35_{\pm0.52}$
+Rand	$44.59_{\pm1.15}$	$70.55_{\pm0.42}$	$63.31_{\pm1.74}$
+SSM	$50.51_{\pm1.15}$	$69.31_{\pm0.77}$	$64.41_{\pm1.04}$
+Meta-learned-ROPES	$53.71_{\pm2.33}$	$73.05_{\pm0.98}$	$62.93_{\pm1.28}$
+Meta-learned-WIQA	$48.30_{\pm0.69}$	$72.38_{\pm0.37}$	$63.14_{\pm1.26}$
+Meta-learned-QuaRTz	$49.01_{\pm1.92}$	$72.65_{\pm0.53}$	$63.69_{\pm0.48}$

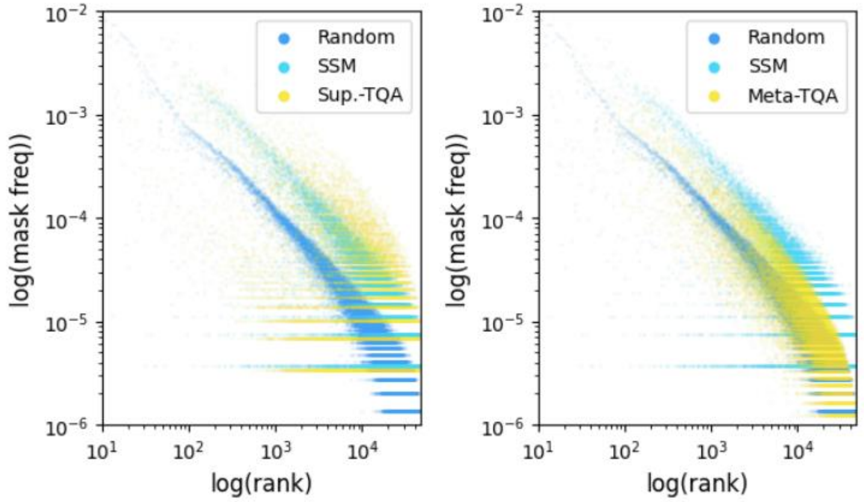
Analysis – Quantitative Analysis

Relation to Part-of-speech Tags



The learned masks are more ***customized*** towards the downstream tasks.

Relation to Token Frequency



(a) Supervised-TriviaQA (b) Meta-learned-TriviaQA

Random → approximates a Zipfian Distribution
SSM → Weaker preference for Zipfian
Learned Policies → Even weaker preference

Conclusions

- We introduce ***an analysis protocol*** to study the influence brought by different masking policies.
- We describe ***two methods (supervised/meta-learned masking policies)*** to learn masking policies.
- We conduct a large-scale ***empirical study*** with ***9 NLP tasks*** and ***3 categories of masking policies***.
- We ...
    - identify several ***successful cases*** of intermediate pre-training,
    - offer ***in-depth analysis and insights*** for the masking policies we used,
    - discuss the ***pros and cons*** of learned masking policies,
    - ***summarize several suggestions and tips*** for researchers who wish to adopt intermediate pre-training in their applications.