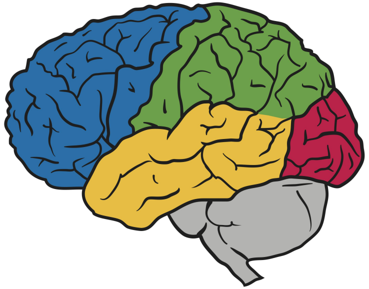
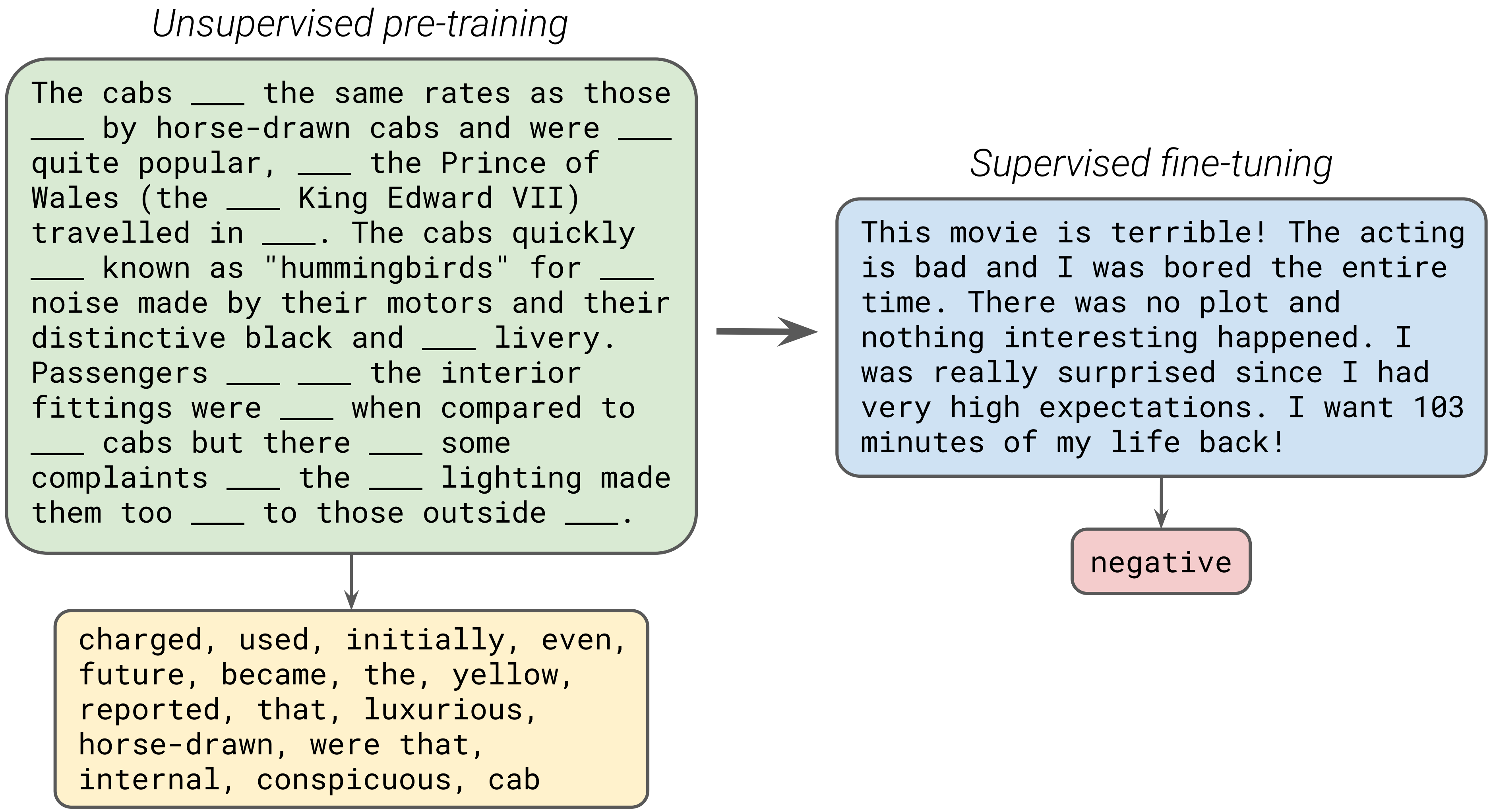
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

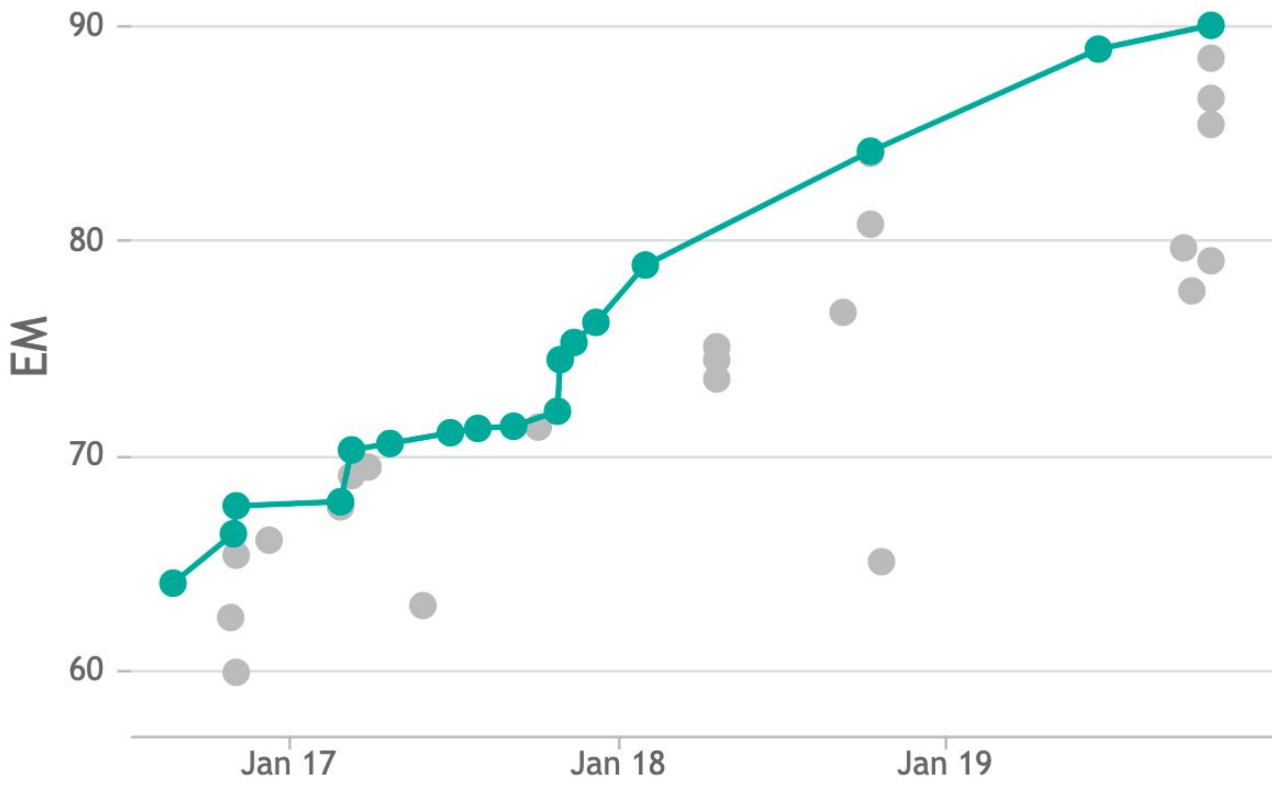
Colin Raffel

with Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu

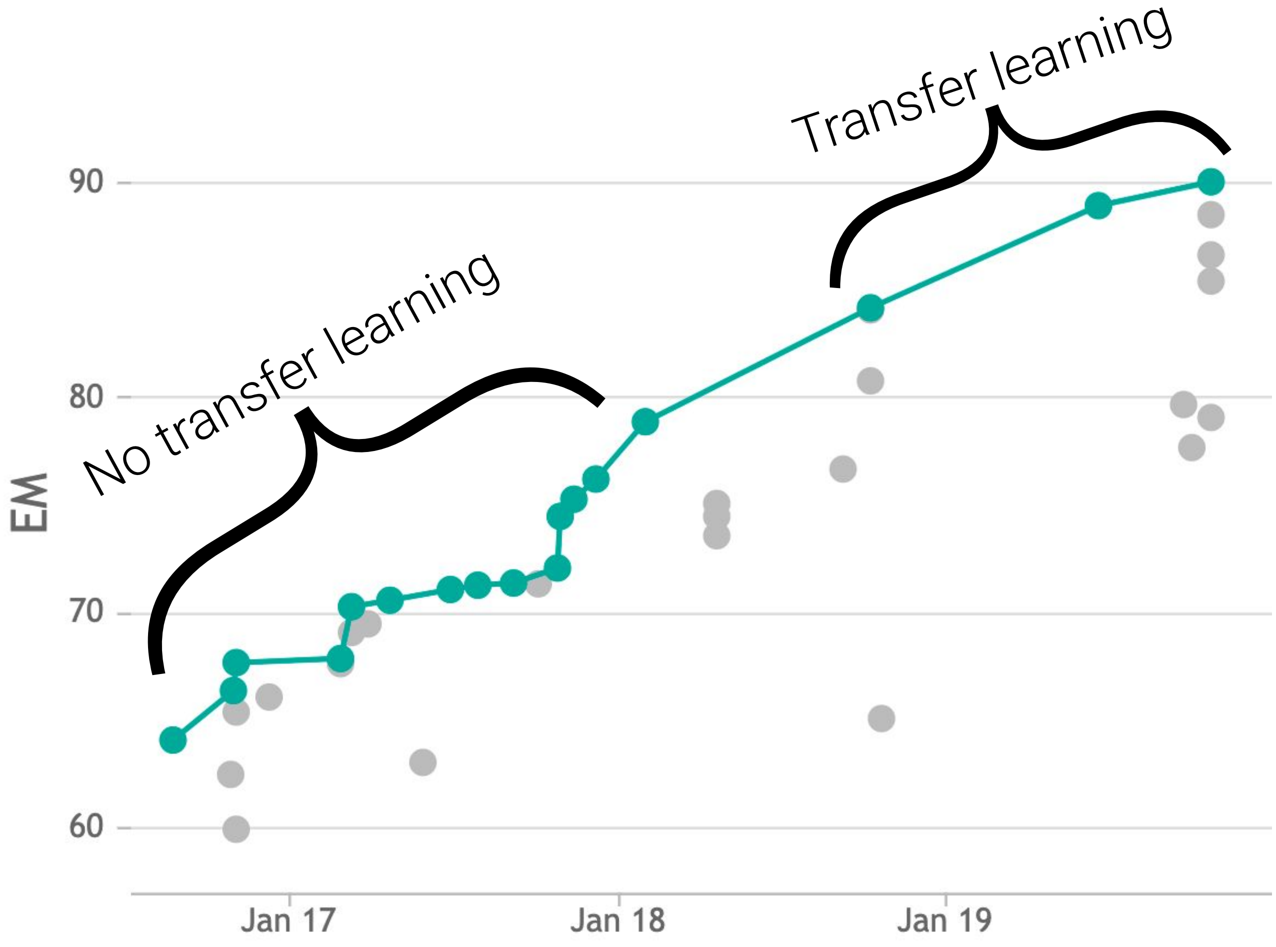



Unsupervised pre-training
The cabs ___ the same rates as those ___ by horse-drawn cabs and were ___ quite popular, ___ the Prince of Wales (the ___ King Edward VII) travelled in ___. The cabs quickly ___ known as "hummingbirds" for ___ noise made by their motors and their distinctive black and ___ livery. Passengers ___ ___ the interior fittings were ___ when compared to ___ cabs but there ___ some complaints ___ the ___ lighting made them too ___ to those outside ___.
charged, used, initially, even, future, became, the, yellow, reported, that, luxurious, horse-drawn, were that, internal, conspicuous, cab
Supervised fine-tuning
This movie is terrible! The acting is bad and I was bored the entire time. There was no plot and nothing interesting happened. I was really surprised since I had very high expectations. I want 103 minutes of my life back!
negative

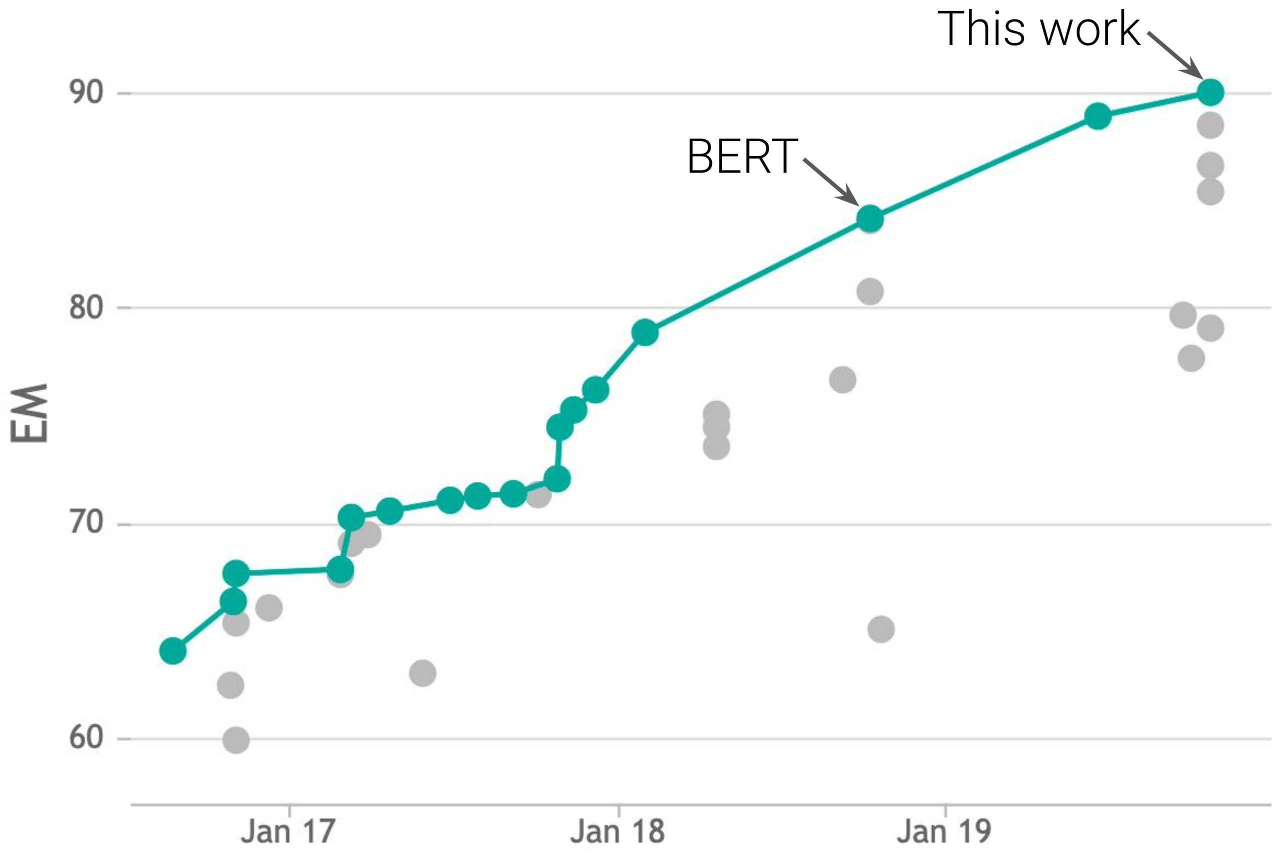
SQuAD Exact Match score (validation set)



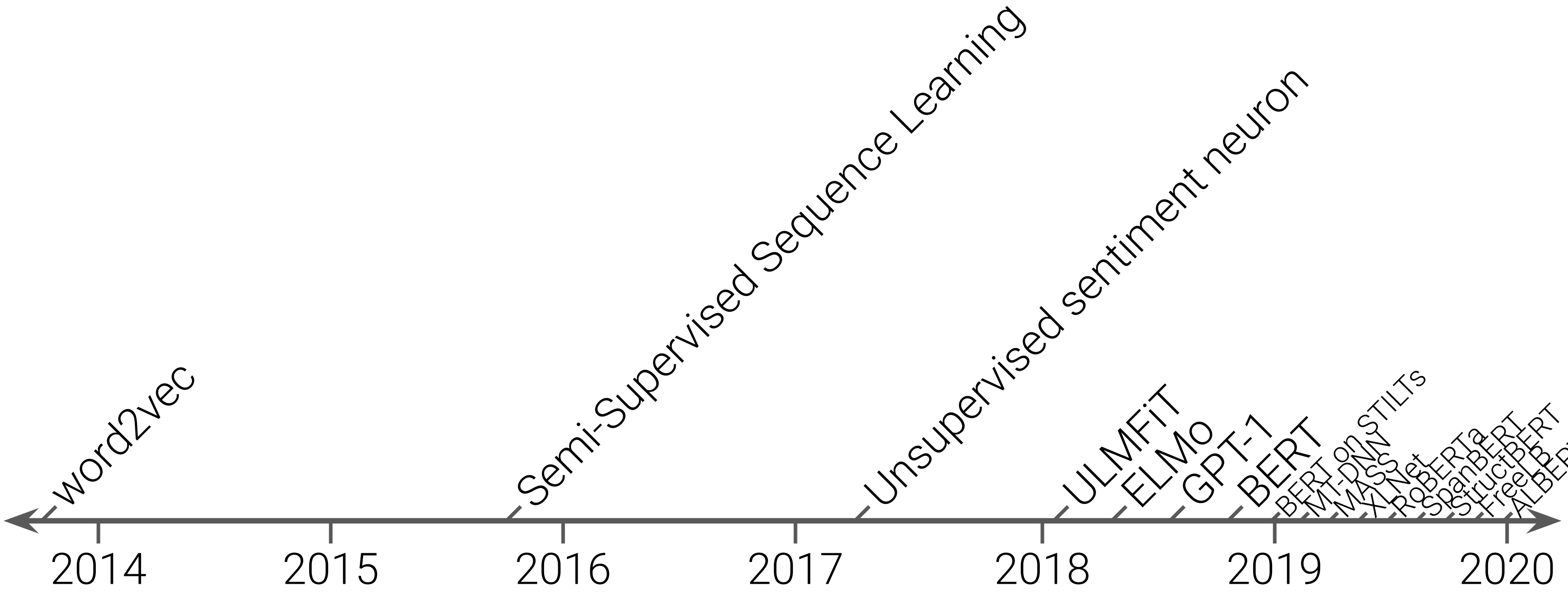
Source: https://paperswithcode.com/sota/question-answering-on-squad11-dev



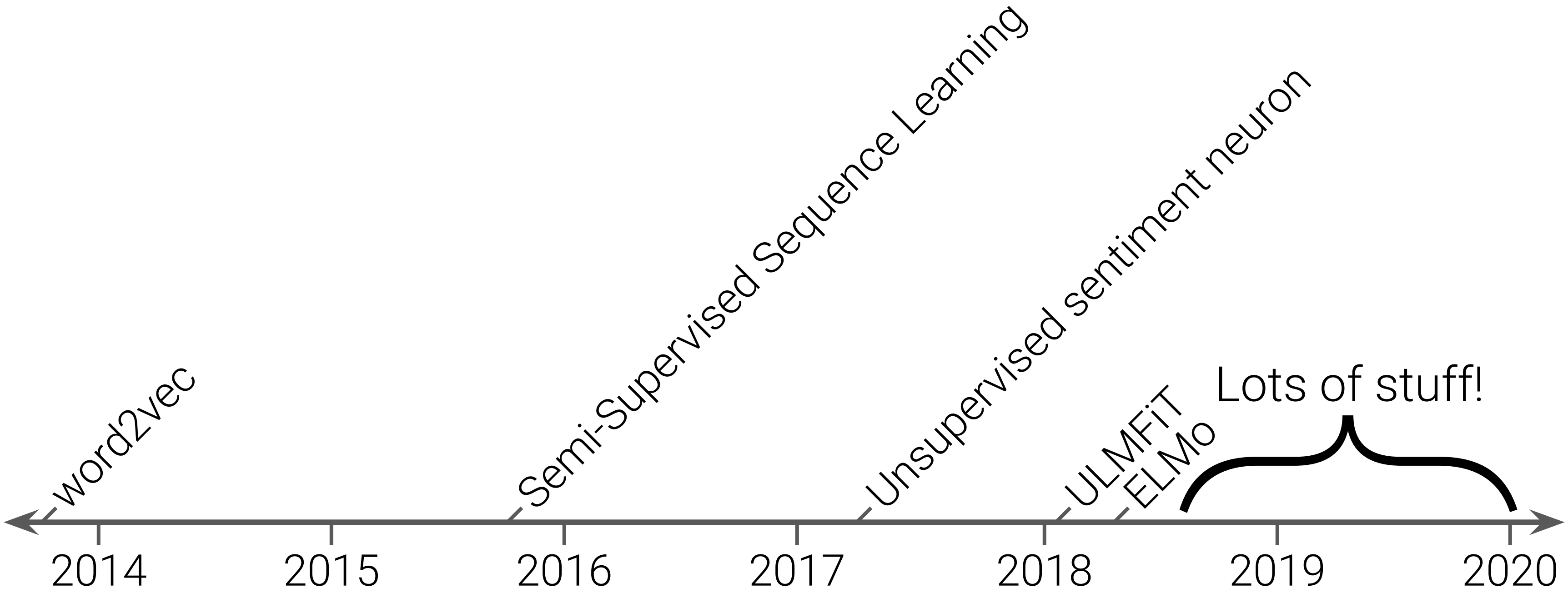
Source: https://paperswithcode.com/sota/question-answering-on-squad11-dev



Source: https://paperswithcode.com/sota/question-answering-on-squad11-dev



word2vec
Semi-Supervised Sequence Learning
Unsupervised sentiment neuron
ULMFiT
ELMo
GPT-1
BERT
BERT on STILTs
MT-DNN
MASS
XLNet
RoBERTa
SpanBERT
StructBERT
FreeLB
ALBERT
2014
2015
2016
2017
2018
2019
2020


word2vec
Semi-Supervised Sequence Learning
Unsupervised sentiment neuron
ULMFiT
ELMo
Lots of stuff!
2014
2015
2016
2017
2018
2019
2020

- Paper A proposes an unsupervised pre-training technique called "FancyLearn".
- Paper B proposes another pre-training technique called "FancierLearn" and achieves better results.
- Paper A uses **Wikipedia** for unlabeled data.
- Paper B uses **Wikipedia and the Toronto Books Corpus**.
- *Is FancierLearn better than FancyLearn?*

- Paper A proposes an unsupervised pre-training technique called "FancyLearn".
- Paper B proposes another pre-training technique called "FancierLearn" and achieves better results.
- Paper A uses a model with **100 million parameters**.
- Paper B uses a model with **200 million parameters**.
- *Is FancierLearn better than FancyLearn?*

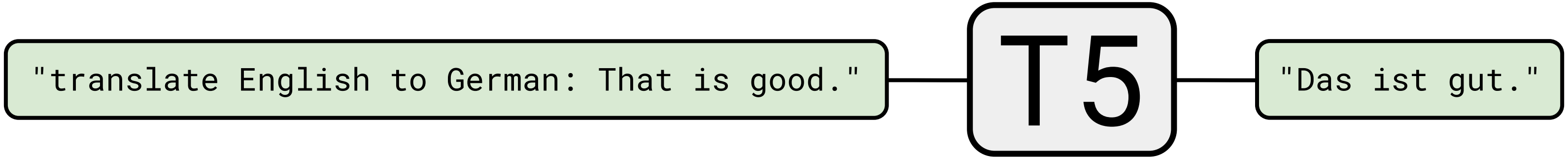
- Paper A proposes an unsupervised pre-training technique called "FancyLearn".
- Paper B proposes another pre-training technique called "FancierLearn" and achieves better results.
- Paper A pre-trains on **100 billion tokens** of unlabeled data.
- Paper B pre-trains on **200 billion tokens** of unlabeled data.
- *Is FancierLearn better than FancyLearn?*

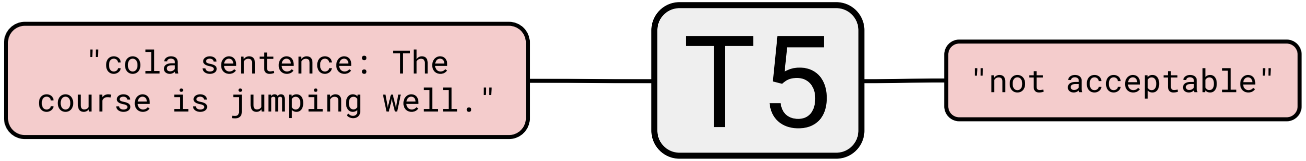
- Paper A proposes an unsupervised pre-training technique called "FancyLearn".
- Paper B proposes another pre-training technique called "FancierLearn" and achieves better results.
- Paper A uses the **Adam optimizer**.
- Paper B uses **SGD with momentum**.
- *Is FancierLearn better than FancyLearn?*

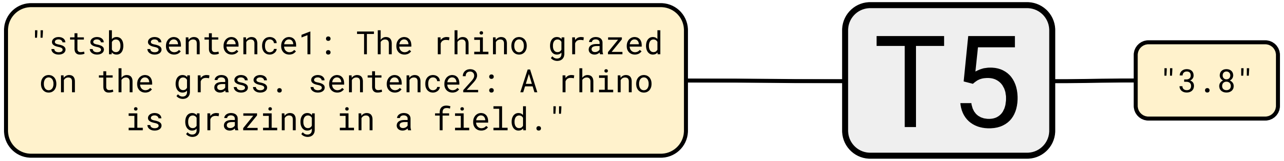
Given the current landscape of transfer learning for NLP, *what works best*? And how far can we push the tools we already have?

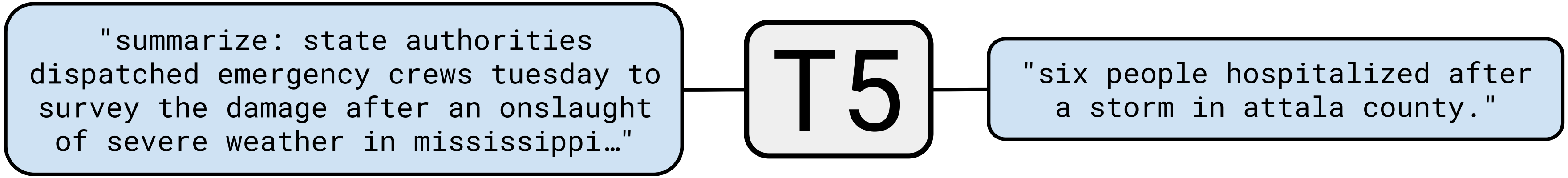
Text-to-Text
Transfer
Transformer

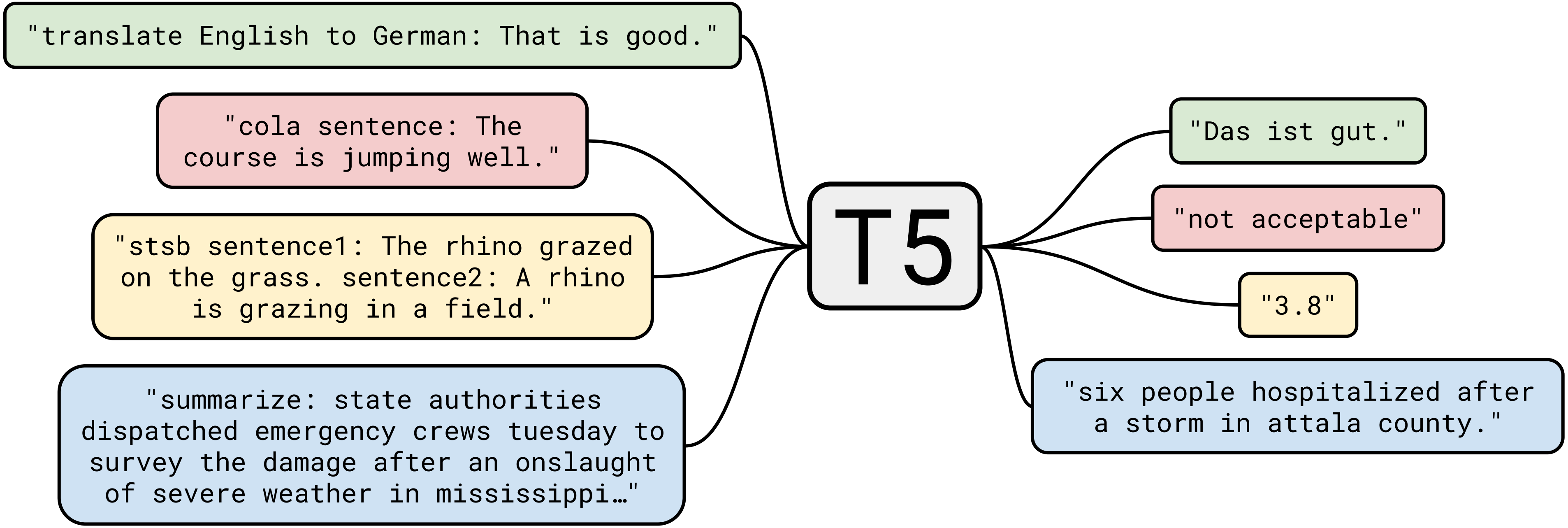
T5

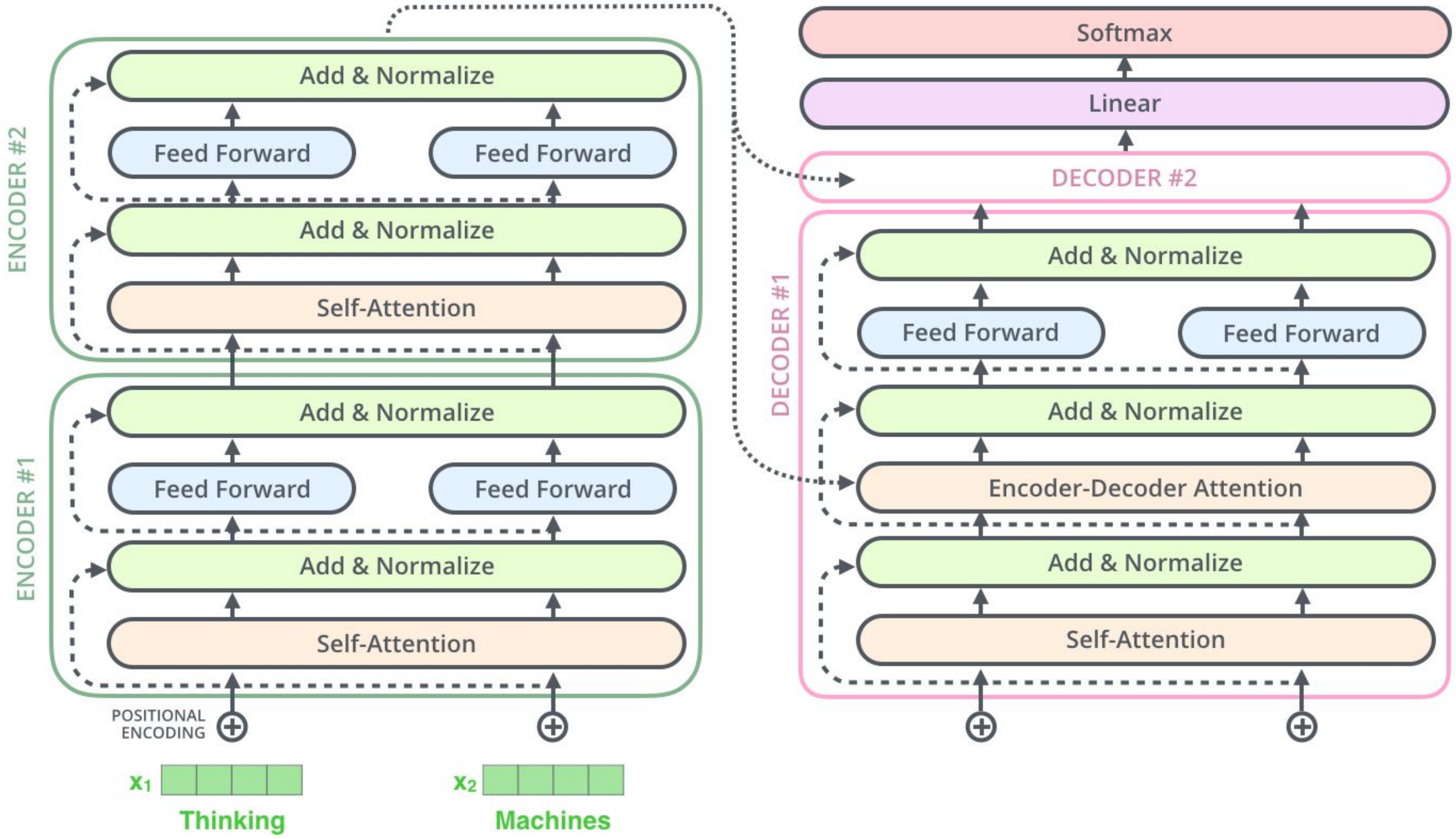

"translate English to German: That is good."
T5
"Das ist gut."


"cola sentence: The course is jumping well."
T5
"not acceptable"


"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
T5
"3.8"


"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"six people hospitalized after a storm in attala county."


"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."



Source: http://jalammar.github.io/illustrated-transformer/

/ookləˌhoʊmə -/), often sho
pital and largest city of the u
ma. the county seat of oklaho
ity ranks 27th among united
tion. the population grew foll
s, with the population estima
ed to 643,648 as of july 2017
oklahoma city metropolitan a
on of 1,358,452,[9] and the
shawnee combined statistica
on of 1,459,758 residents,[9]
oma's largest metropolitan a

city limits extend into canad

running man was cl
variety"; a genre of
environment.[1] the
complete missions
race.[2] the show ha
familiar reality-varie
games. it has garne
comeback program
of the program, afte
family outing in feb

the show has become popu
asia, and has gained online
hallyu fans, having been fansubbed into various
languages, such as english, spanish, portuguese,
french, italian, thai, vietnamese, chinese, ...

county,[8] the ci
cities in populat
the 2010 census
to have increase
as of 2015, the
had a populatio
oklahoma city-s
had a populatio
making it oklah

oklahoma city's

the signing of the treaty formally ended the seven years' war, known as the french and indian war in the north american theatre,[1] and marked the beginning of an era of british dominance outside europe.[2] great britain and france each returned much of the territory that they had captured during the war, but great britain gained much of france's possessions in north america. additionally, great britain agreed to protect roman catholicism in the new world...

s a small hand-propelled vehicle,
one wheel, designed to be
ed by a single person using two
ar, or by a sail to push the
rrow by wind. the term
made of two words: "wheel" and
" is a derivation of the old
which was a device used for

is designed to distribute the
weight of its load between the wheel and the
operator, so enabling the convenient carriage of
heavier and bulkier loads than would be possible
were the weight carried entirely by the operator.
as such it is a second-class lever...

== treaty o

were the weight ca
as such it is a second-class lever...

reaty of paris, also kn
3, was signed on 10 fe
doms of great britain,
ugal in agreement, aft
france a

== lemon

the lemon, citrus limon (l.) osbeck, is a species of
plant family
y north

== wheelbarrow

a wheelbarrow is a small hand-propelled vehicle, usually with just one wheel, designed to be pushed and guided by a single person using two handles at the rear, or by a sail to push the ancient wheelbarrow by wind. the term "wheelbarrow" is made of two words: "wheel" and "barrow." "barrow" is a derivation of the old english "bearwe" which was a device used for carrying loads.

== piano

the piano is an acoustic, stringed musical instrument invented in italy by bartolomeo cristofori around the year 1700 (the exact year is uncertain), in which the strings are struck by hammers. it is played using a keyboard,[1] which is a row of keys (small levers) that the performer presses down or strikes with the fingers and thumbs of both hands to cause the hammers to strike the strings.

the word piano is a shortened form of pianoforte, the italian term for the early 1700s versions of the instrument, which in turn derives from gravicembalo col piano e forte[2] and fortepiano. the italian musical terms piano and forte indicate

paris, also known as the treaty
gned on 10 february 1763 by th
great britain, france and spain,
greement, after great britain's v
nd spain during the seven year

f the treaty formally ended the
nown as the french and indian w
erican theatre,[1] and marked t
an era of british dominance ou
eat britain and france each retu
territory that they had captured
ar, but great britain gained muc
sessions in north america.

Common Crawl Web Extracted Text

Menu

Lemon

Introduction

The lemon, Citrus Limon (l.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae.
The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses.
The juice of the lemon is about 5% to 6% citric acid, with a ph of around 2.2, giving it a sour taste.

Article

The origin of the lemon is unknown, though lemons are thought to have first grown in Assam (a region in northeast India), northern Burma or China.
A genomic study of the lemon indicated it was a hybrid between bitter orange (sour orange) and citron.

Please enable JavaScript to use our site.

Home
Products
Shipping
Contact
FAQ

Dried Lemons, $3.59/pound

Organic dried lemons from our farm in California.
Lemons are harvested and sun-dried for maximum flavor.
Good in soups and on popcorn.

The lemon, Citrus Limon (l.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae.
The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses.
The juice of the lemon is about 5% to 6% citric acid, with a ph of around 2.2, giving it a sour taste.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Curabitur in tempus quam. In mollis et ante at consectetur.
Aliquam erat volutpat.
Donec at lacinia est.
Duis semper, magna tempor interdum suscipit, ante elit molestie urna, eget efficitur risus nunc ac elit.
Fusce quis blandit lectus.
Mauris at mauris a turpis tristique lacinia at nec ante.
Aenean in scelerisque tellus, a efficitur ipsum.
Integer justo enim, ornare vitae sem non, mollis fermentum lectus.
Mauris ultrices nisl at libero porta sodales in ac orci.

```
function Ball(r) {
   this.radius = r;
   this.area = pi * r ** 2;
   this.show = function(){
      drawCircle(r);
   }
}
```

Common Crawl Web Extracted Text

Menu

Lemon

Introduction

The lemon, Citrus Limon (l.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae.
The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses.
The juice of the lemon is about 5% to 6% citric acid, with a ph of around 2.2, giving it a sour taste.

Article

The origin of the lemon is unknown, though lemons are thought to have first grown in Assam (a region in northeast India), northern Burma or China.
A genomic study of the lemon indicated it was a hybrid between bitter orange (sour orange) and citron.

Please enable JavaScript to use our site.

Home
Products
Shipping
Contact
FAQ

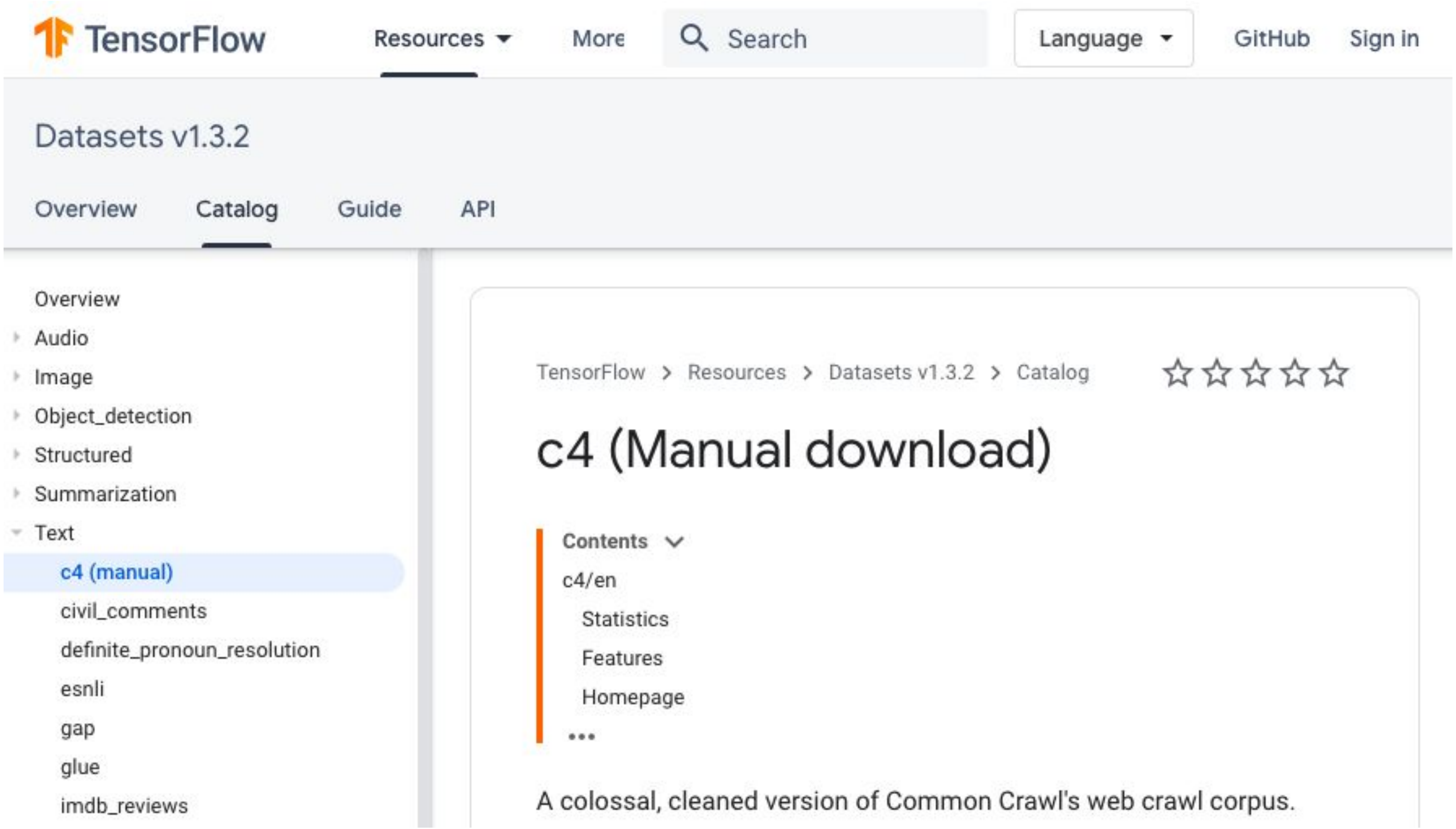
Dried Lemons, $3.59/pound

Organic dried lemons from our farm in California.
Lemons are harvested and sun-dried for maximum flavor.
Good in soups and on popcorn.

The lemon, Citrus Limon (l.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae.
The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses.
The juice of the lemon is about 5% to 6% citric acid, with a ph of around 2.2, giving it a sour taste.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Curabitur in tempus quam. In mollis et ante at consectetur.
Aliquam erat volutpat.
Donec at lacinia est.
Duis semper, magna tempor interdum suscipit, ante elit molestie urna, eget efficitur risus nunc ac elit.
Fusce quis blandit lectus.
Mauris at mauris a turpis tristique lacinia at nec ante.
Aenean in scelerisque tellus, a efficitur ipsum.
Integer justo enim, ornare vitae sem non, mollis fermentum lectus.
Mauris ultrices nisl at libero porta sodales in ac orci.

```
function Ball(r) {
  this.radius = r;
  this.area = pi * r ** 2;
  this.show = function(){
    drawCircle(r);
  }
}
```

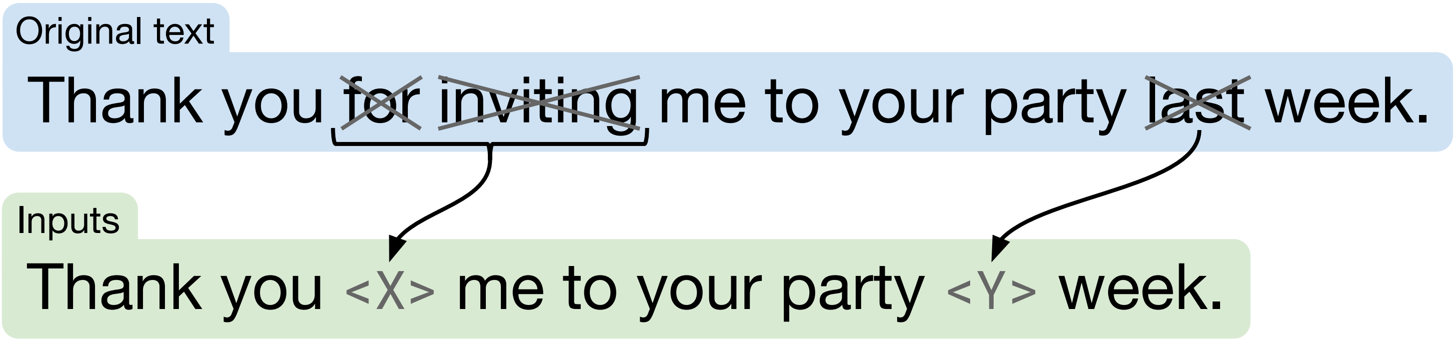

TensorFlow
Resources
More
Search
Language
GitHub
Sign in
Datasets v1.3.2
Overview
Catalog
Guide
API
Overview
Audio
Image
Object_detection
Structured
Summarization
Text
c4 (manual)
civil_comments
definite_pronoun_resolution
esnli
gap
glue
imdb_reviews
TensorFlow > Resources > Datasets v1.3.2 > Catalog
c4 (Manual download)
Contents
c4/en
Statistics
Features
Homepage
...
A colossal, cleaned version of Common Crawl's web crawl corpus.

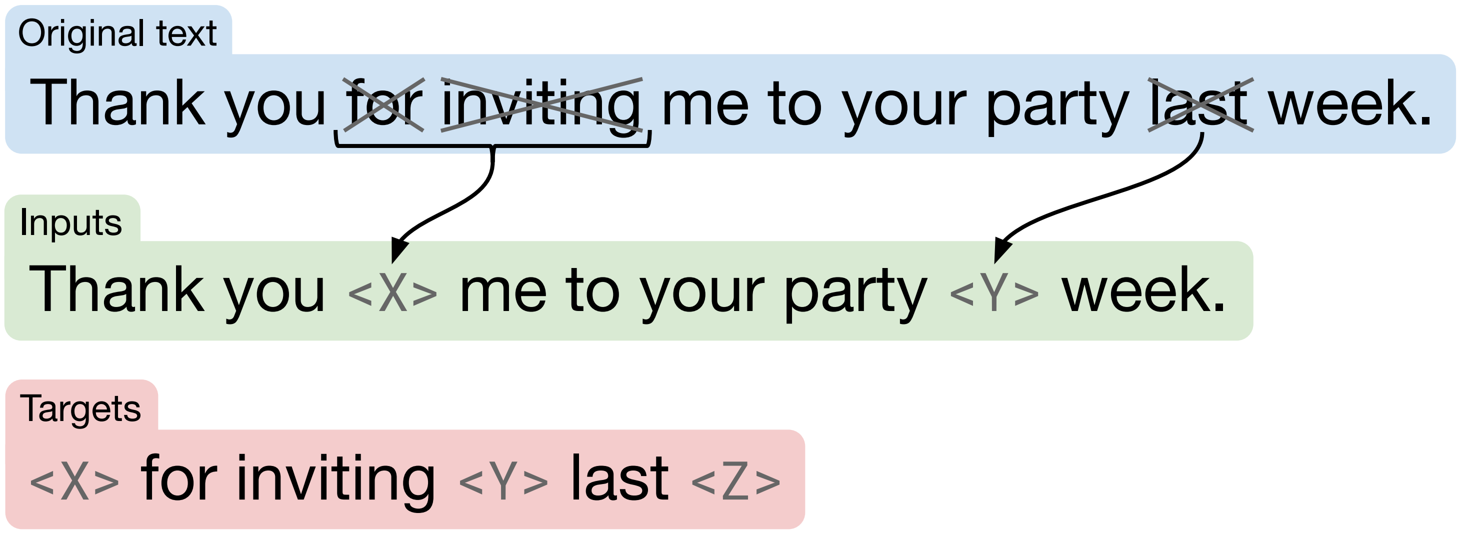
Original text

Thank you for inviting me to your party last week.

Original text

Thank you ~~for~~ ~~inviting~~ me to your party ~~last~~ week.


Original text
Thank you ~~for~~ ~~inviting~~ me to your party ~~last~~ week.
Inputs
Thank you <X> me to your party <Y> week.



Original text
Thank you ~~for inviting~~ me to your party ~~last~~ week.
Inputs
Thank you <X> me to your party <Y> week.
Targets
<X> for inviting <Y> last <Z>

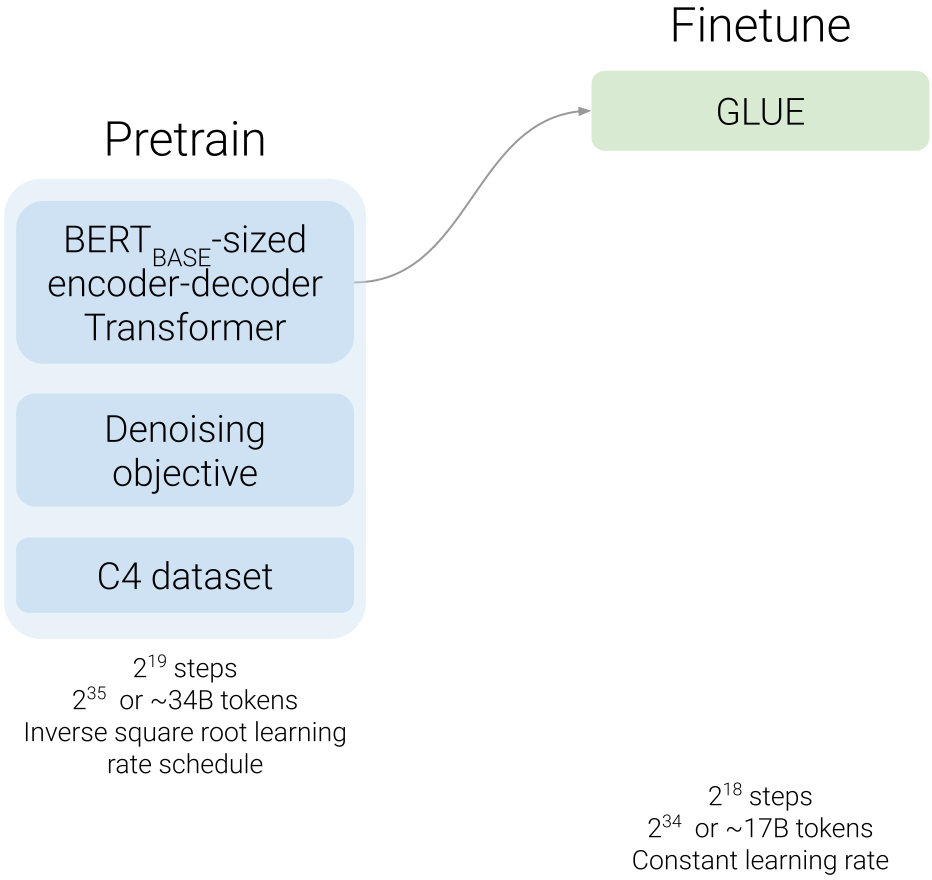
Pretrain

BERT$_{BASE}$-sized encoder-decoder Transformer

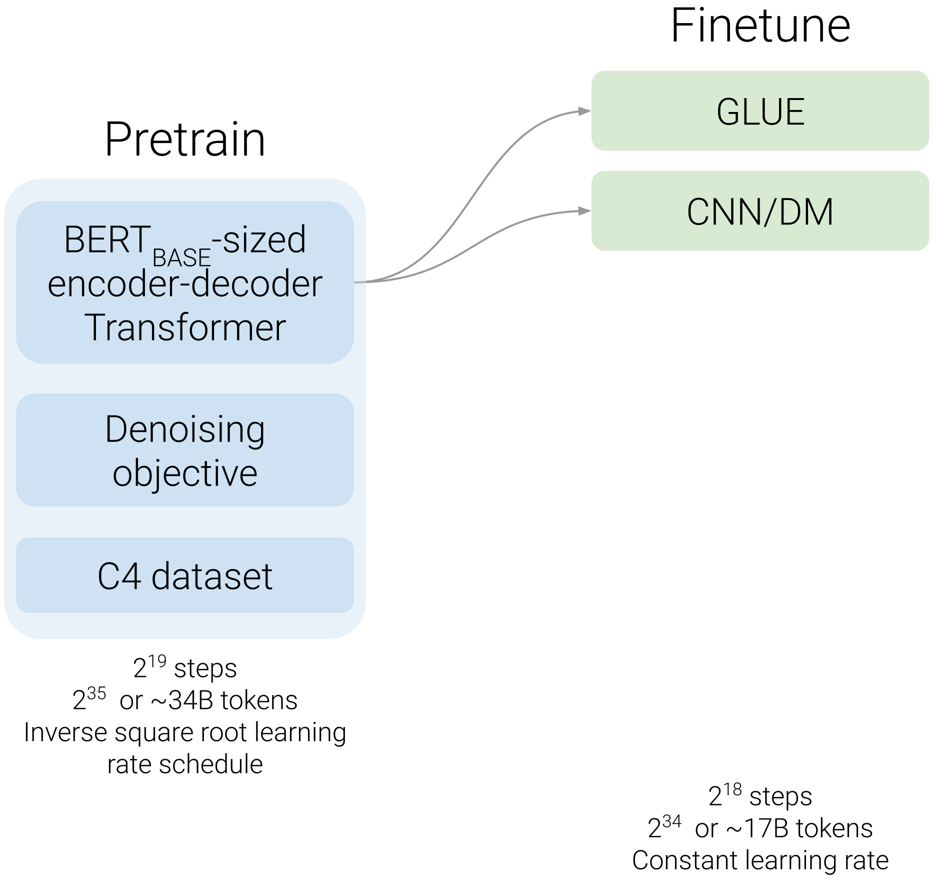
Denoising objective

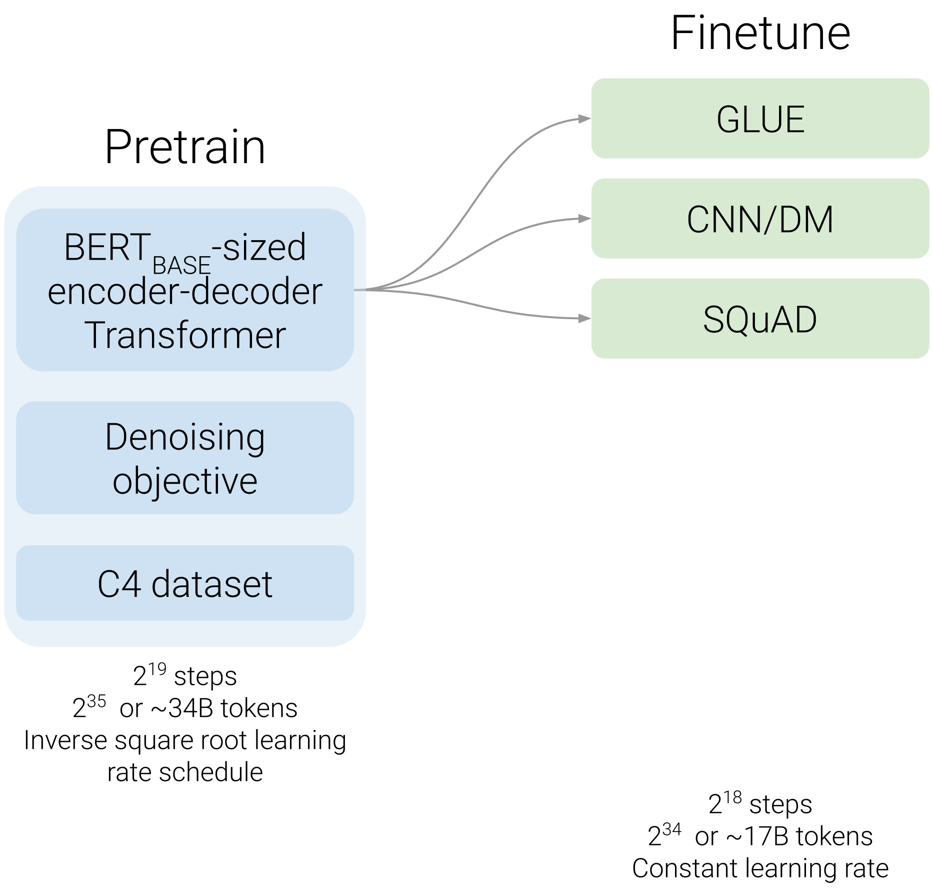
C4 dataset

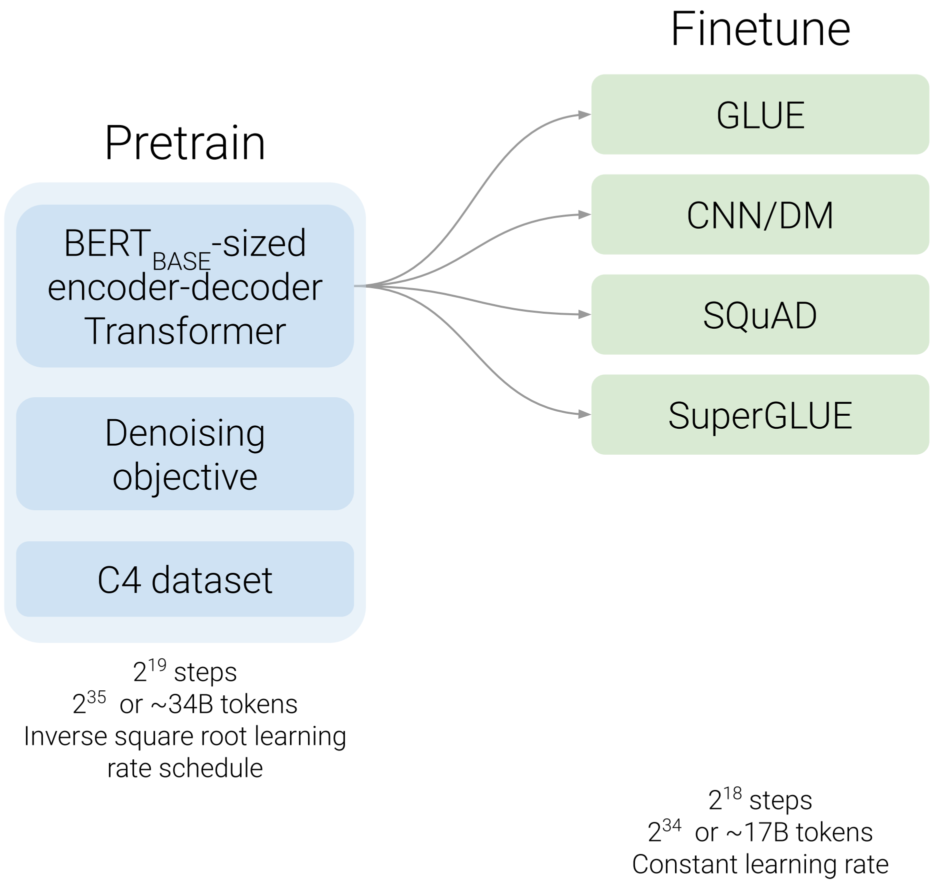
2^{19} steps
2^{35} or ~34B tokens
Inverse square root learning rate schedule

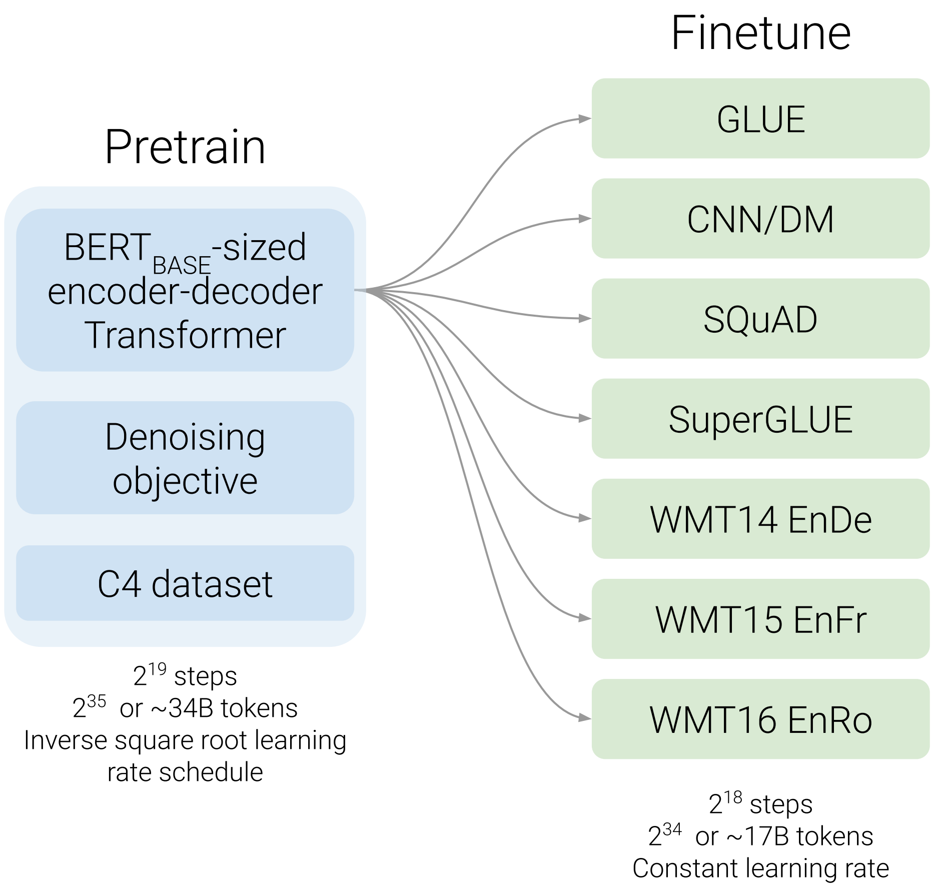


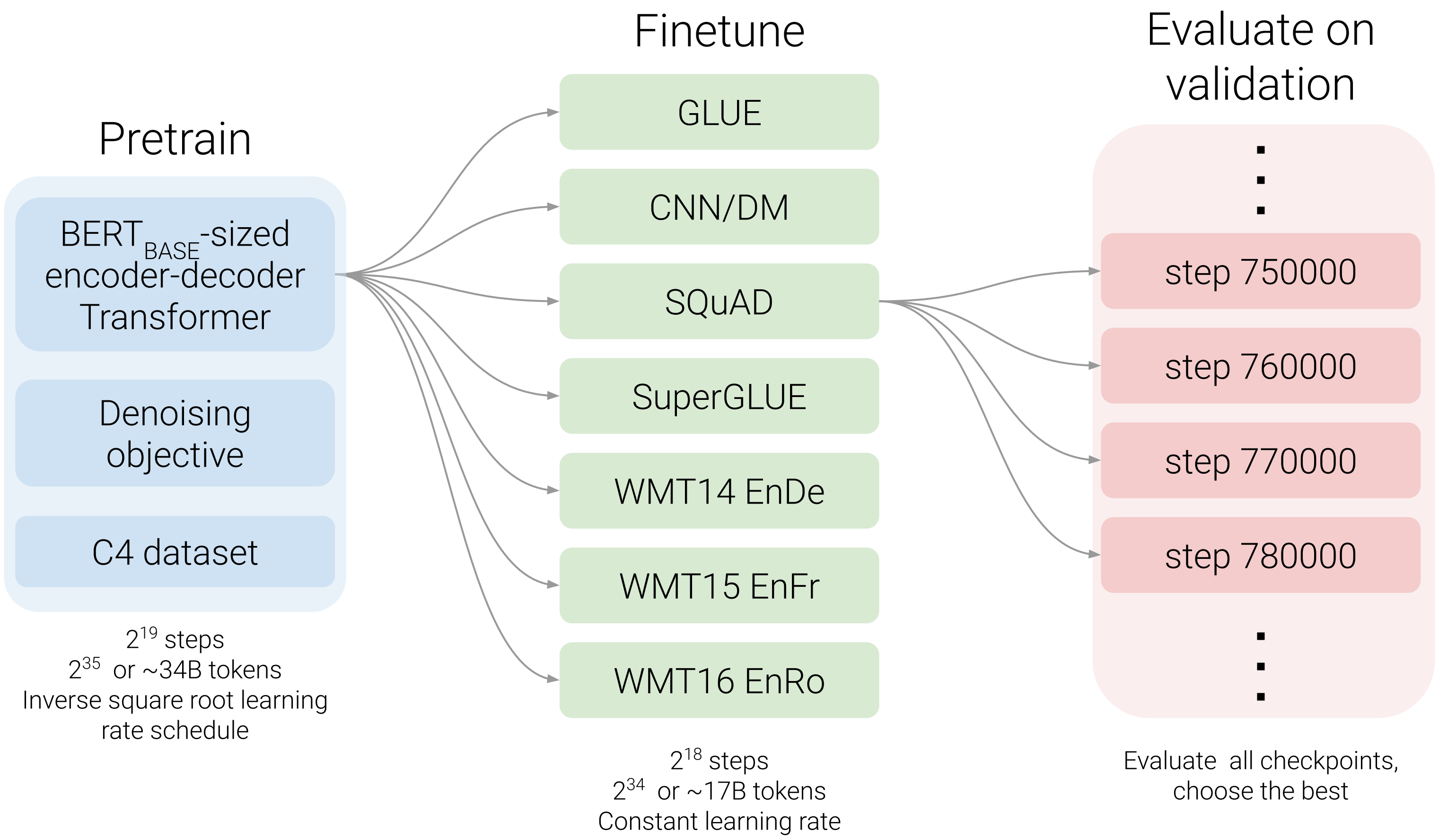
Finetune
GLUE
Pretrain
BERT$_{BASE}$-sized encoder-decoder Transformer
Denoising objective
C4 dataset
2^{19} steps
2^{35} or ~34B tokens
Inverse square root learning rate schedule
2^{18} steps
2^{34} or ~17B tokens
Constant learning rate


Pretrain
BERT$_{BASE}$-sized encoder-decoder Transformer
Denoising objective
C4 dataset
2^{19} steps
2^{35} or ~34B tokens
Inverse square root learning rate schedule
Finetune
GLUE
CNN/DM
2^{18} steps
2^{34} or ~17B tokens
Constant learning rate


Finetune
GLUE
CNN/DM
SQuAD
Pretrain
BERT BASE-sized encoder-decoder Transformer
Denoising objective
C4 dataset
2^19 steps
2^35 or ~34B tokens
Inverse square root learning rate schedule
2^18 steps
2^34 or ~17B tokens
Constant learning rate


Pretrain
BERTBASE-sized encoder-decoder Transformer
Denoising objective
C4 dataset
219 steps
235 or ~34B tokens
Inverse square root learning rate schedule
Finetune
GLUE
CNN/DM
SQuAD
SuperGLUE
218 steps
234 or ~17B tokens
Constant learning rate


Finetune
Pretrain
BERT_BASE-sized encoder-decoder Transformer
Denoising objective
C4 dataset
GLUE
CNN/DM
SQuAD
SuperGLUE
WMT14 EnDe
WMT15 EnFr
WMT16 EnRo
2^19 steps
2^35 or ~34B tokens
Inverse square root learning rate schedule
2^18 steps
2^34 or ~17B tokens
Constant learning rate



Pretrain
BERT BASE-sized encoder-decoder Transformer
Denoising objective
C4 dataset
2^19 steps
2^35 or ~34B tokens
Inverse square root learning rate schedule
Finetune
GLUE
CNN/DM
SQuAD
SuperGLUE
WMT14 EnDe
WMT15 EnFr
WMT16 EnRo
2^18 steps
2^34 or ~17B tokens
Constant learning rate
Evaluate on validation
step 750000
step 760000
step 770000
step 780000
Evaluate all checkpoints, choose the best

	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Setting 1 *Setting 2* ...	*Downstream task performance*						

	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline average	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Baseline standard deviation	0.235	0.065	0.343	0.416	0.112	0.090	0.108
No pre-training	66.22	17.60	50.31	53.04	25.86	**39.77**	24.04

Star denotes baseline

Comparable to BERT

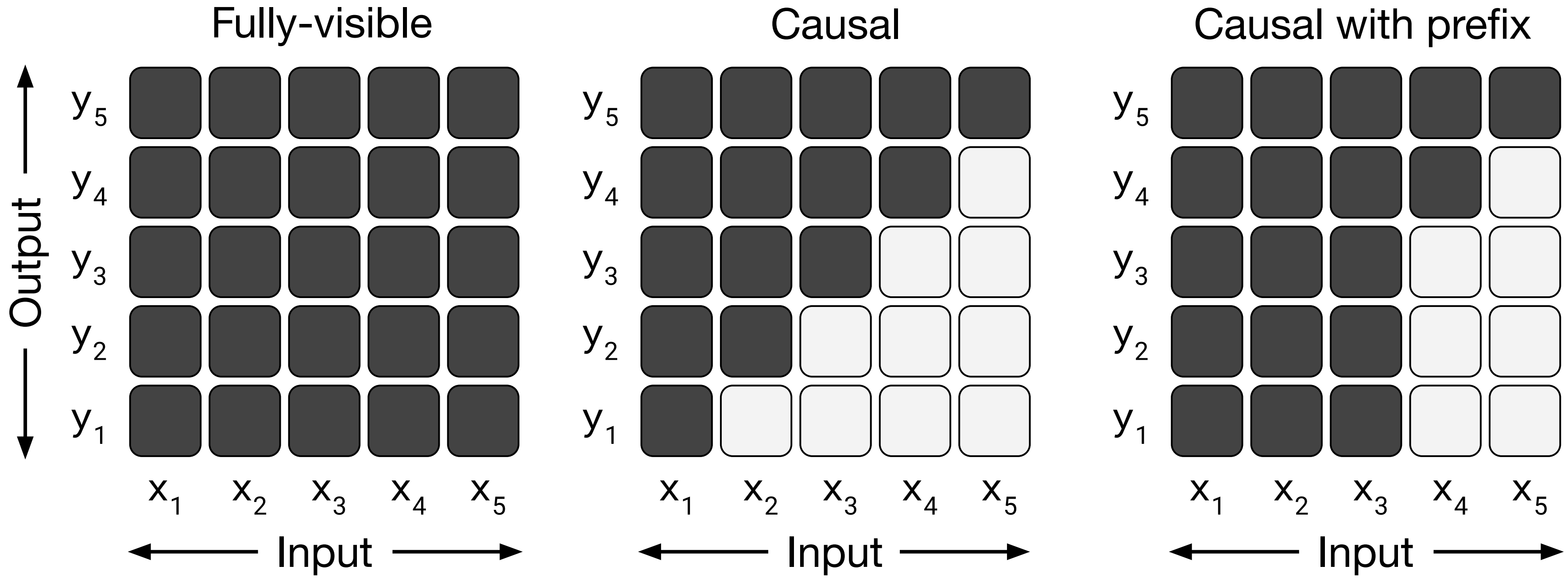
Bold = 1 std. dev. of max

	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline average	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Baseline standard deviation	0.235	0.065	0.343	0.416	0.112	0.090	0.108
No pre-training	66.22	17.60	50.31	53.04	25.86	**39.77**	24.04

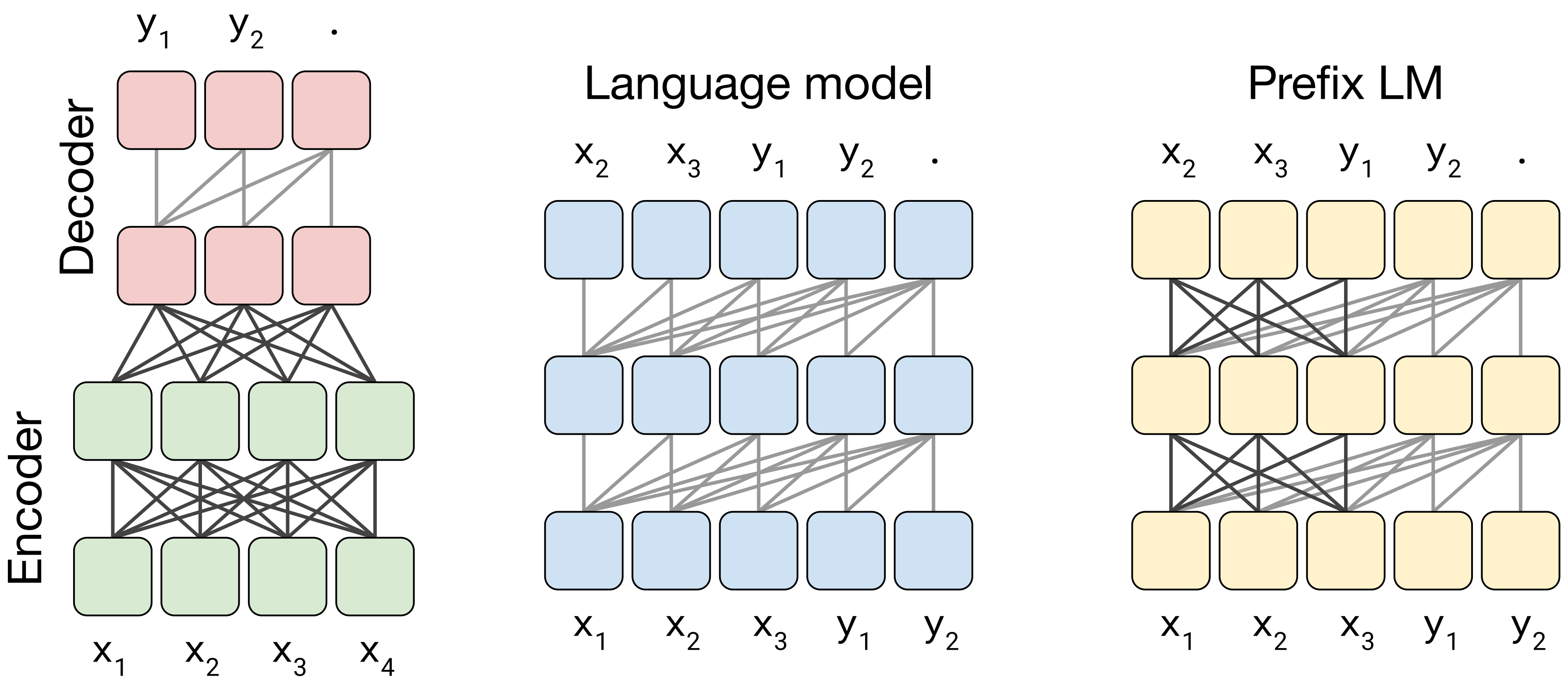
Big training set

Disclaimer

Architectures

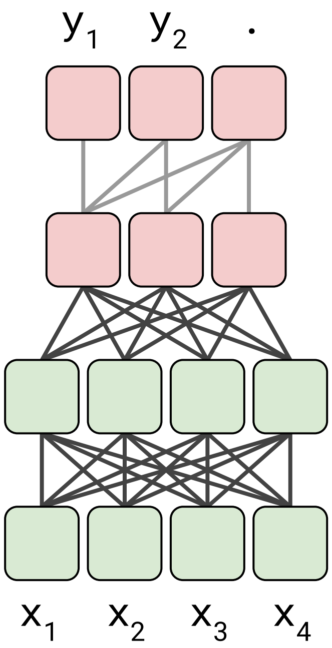


Fully-visible
Causal
Causal with prefix
Output
y_5
y_4
y_3
y_2
y_1
x_1 x_2 x_3 x_4 x_5
Input

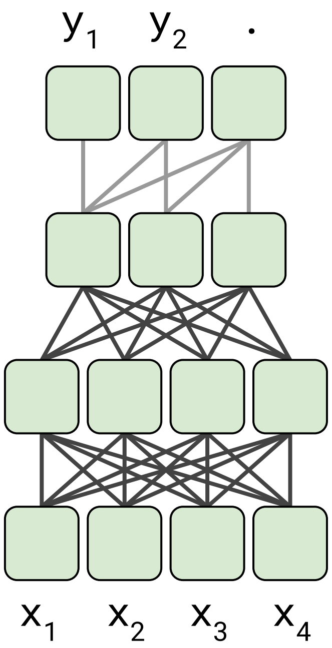


y1 y2 .
Decoder
Encoder
x1 x2 x3 x4
Language model
x2 x3 y1 y2 .
x1 x2 x3 y1 y2
Prefix LM
x2 x3 y1 y2 .
x1 x2 x3 y1 y2

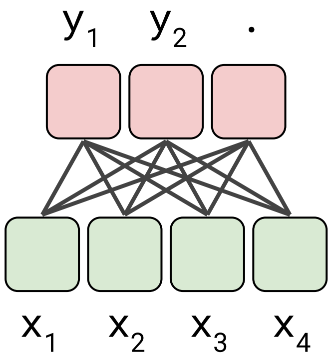
Architecture	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**



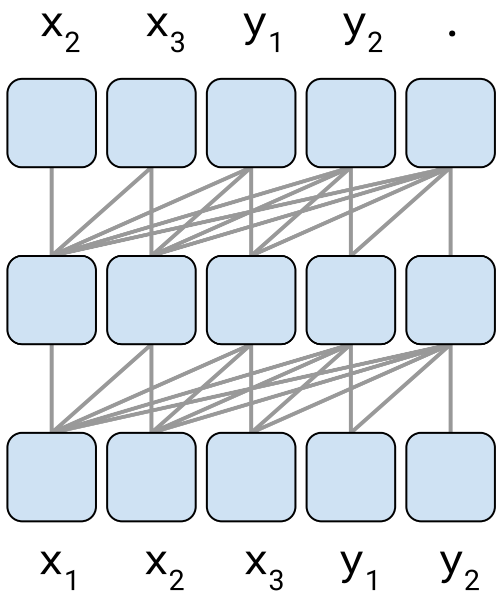
Architecture	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**



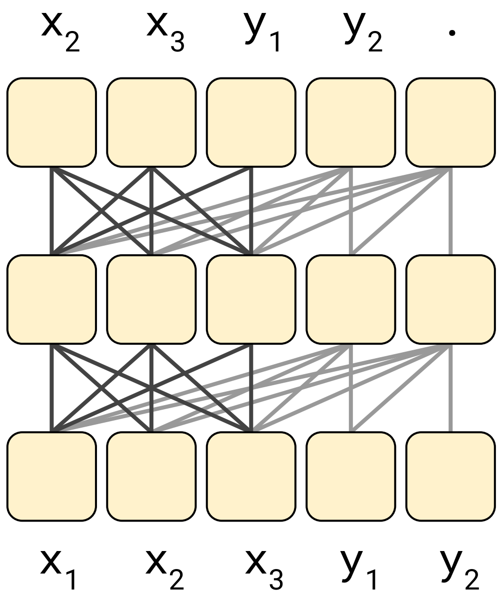
Architecture	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95



Architecture	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86



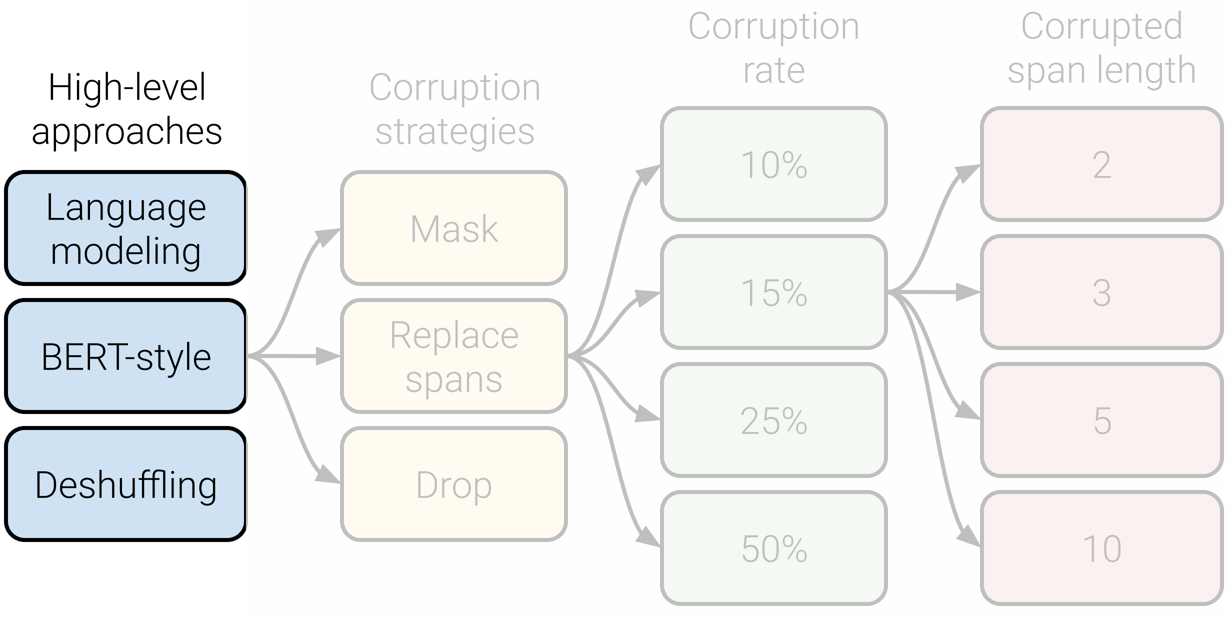
Architecture	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39



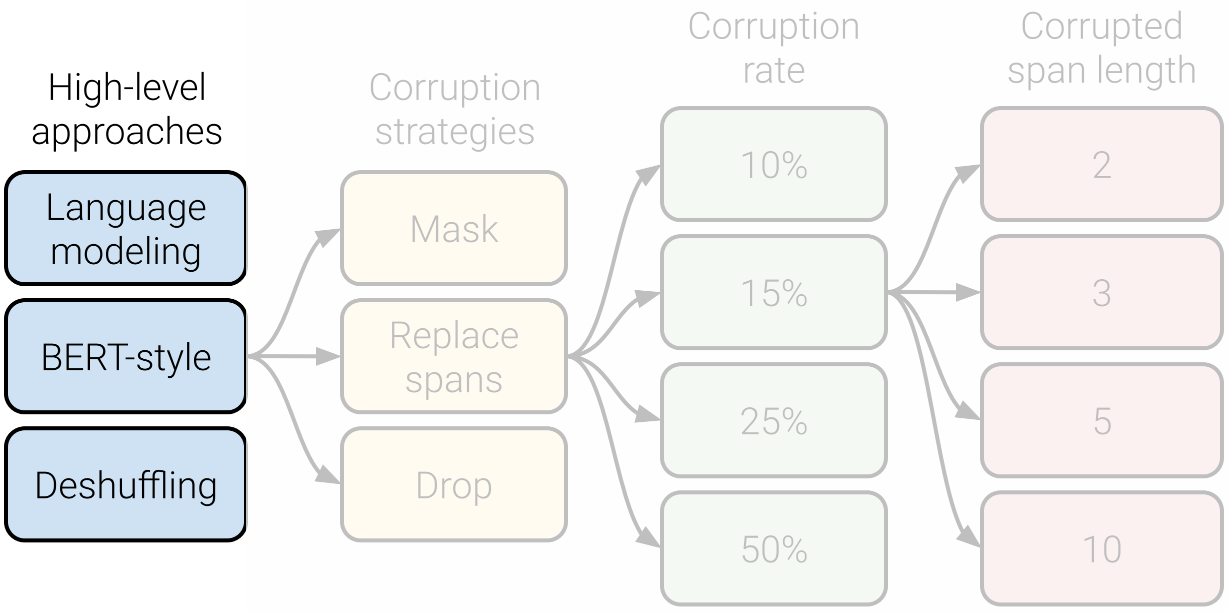
Architecture	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Encoder-decoder	$2P$	M	79.56	18.59	76.02	64.29	26.27	39.17	26.86
Enc-dec, shared	P	M	79.60	18.13	76.35	63.50	26.62	39.17	27.05
Enc-dec, 6 layers	P	$M/2$	78.67	18.26	75.32	64.06	26.13	38.42	26.89
Language model	P	M	73.78	17.54	53.81	56.51	25.23	34.31	25.38
Prefix LM	P	M	79.68	17.84	76.87	64.86	26.28	37.51	26.76

(autoregressive objective)

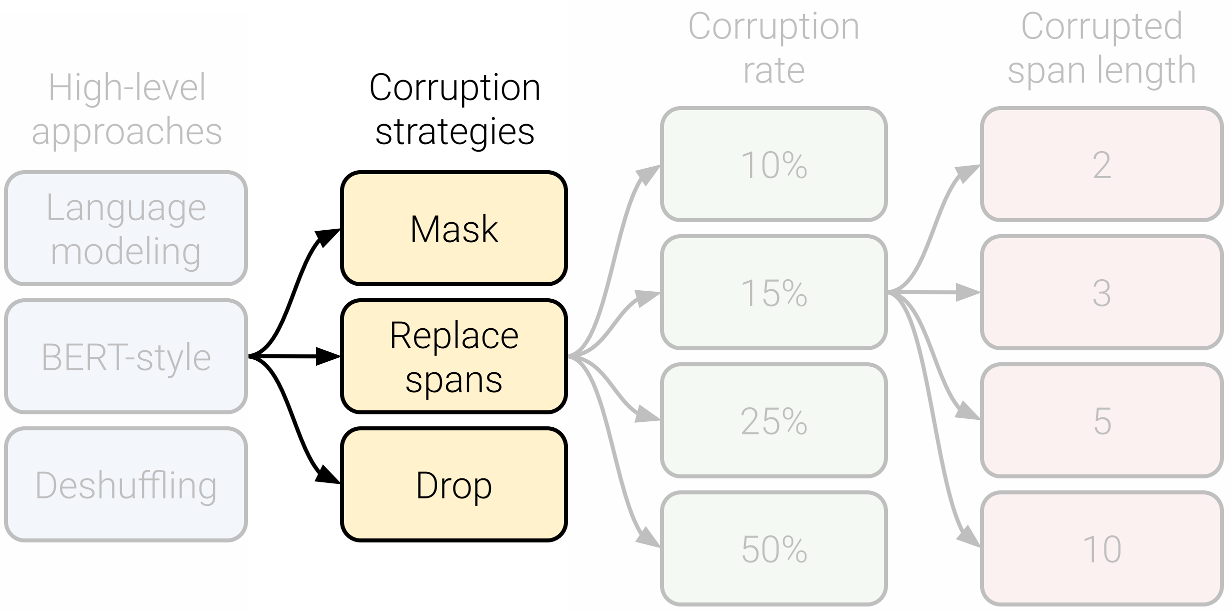
Objectives



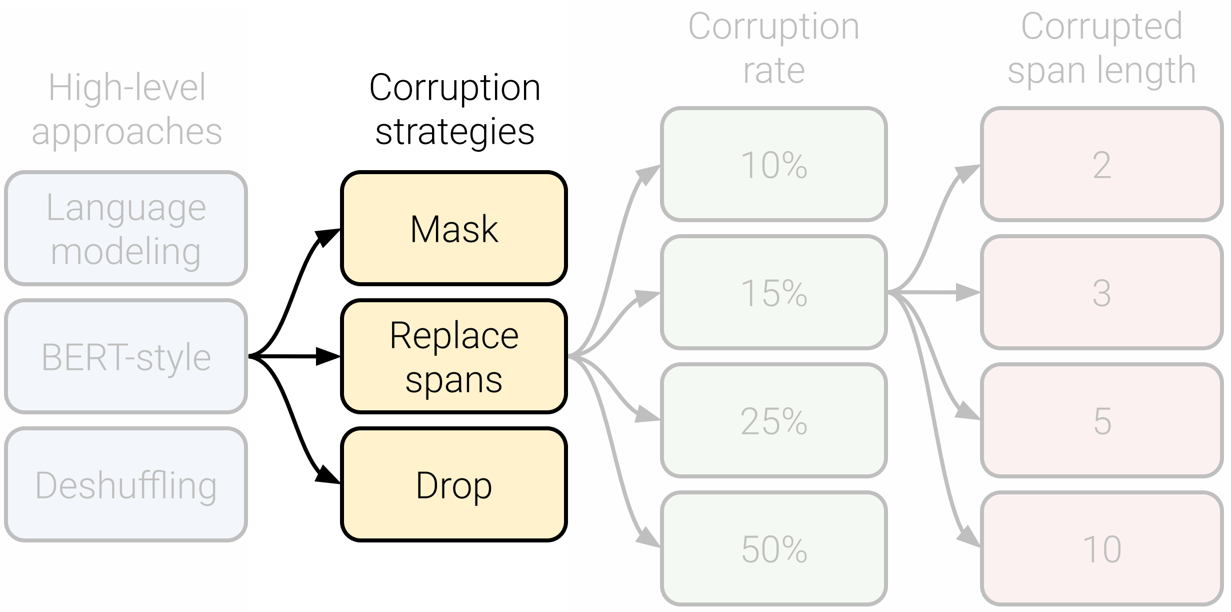
Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
I.i.d. noise, mask tokens	Thank you <M> <M> me to your party <M> week .	*(original text)*
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>



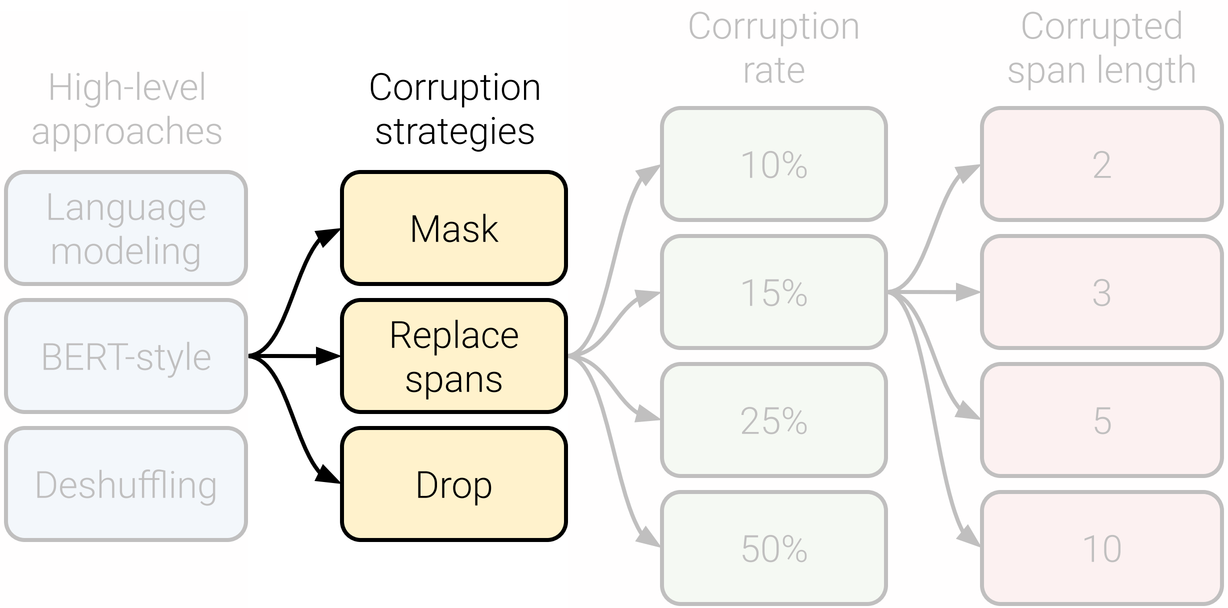
Objective	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Prefix language modeling	80.69	18.94	77.99	65.27	**26.86**	39.73	**27.49**
BERT-style [Devlin et al., 2018]	**82.96**	**19.17**	**80.65**	**69.85**	**26.78**	**40.03**	**27.41**
Deshuffling	73.17	18.59	67.61	58.47	26.11	39.30	25.62



Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
I.i.d. noise, mask tokens	Thank you <M> <M> me to your party <M> week .	*(original text)*
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>



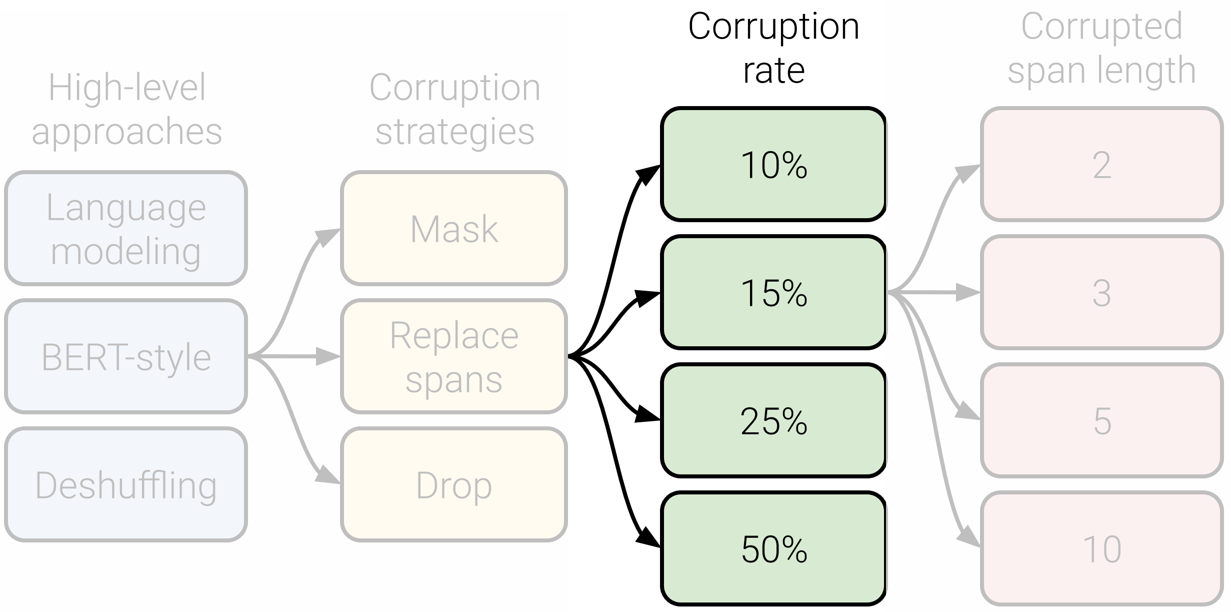
Objective	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
BERT-style [Devlin et al., 2018]	82.96	19.17	**80.65**	69.85	26.78	**40.03**	27.41
MASS-style [Song et al., 2019]	82.32	19.16	80.10	69.28	26.79	**39.89**	27.55
★ Replace corrupted spans	83.28	**19.24**	**80.88**	**71.36**	**26.98**	39.82	**27.65**
Drop corrupted tokens	**84.44**	**19.31**	**80.52**	68.67	**27.07**	39.76	**27.82**



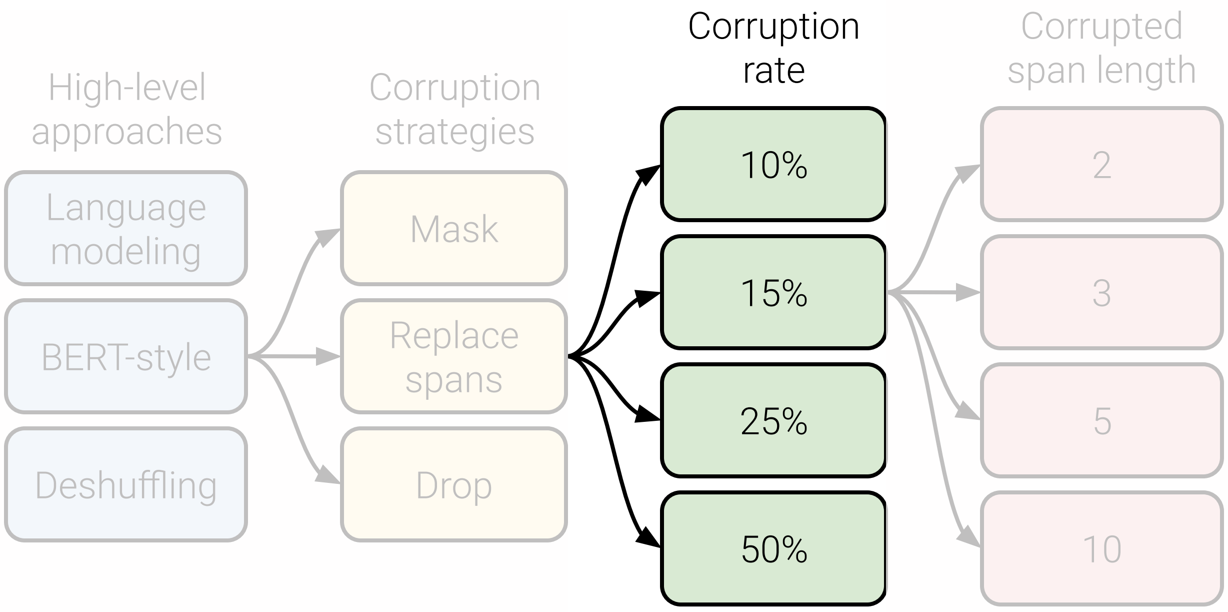
Much better on CoLA

Much better on COPA

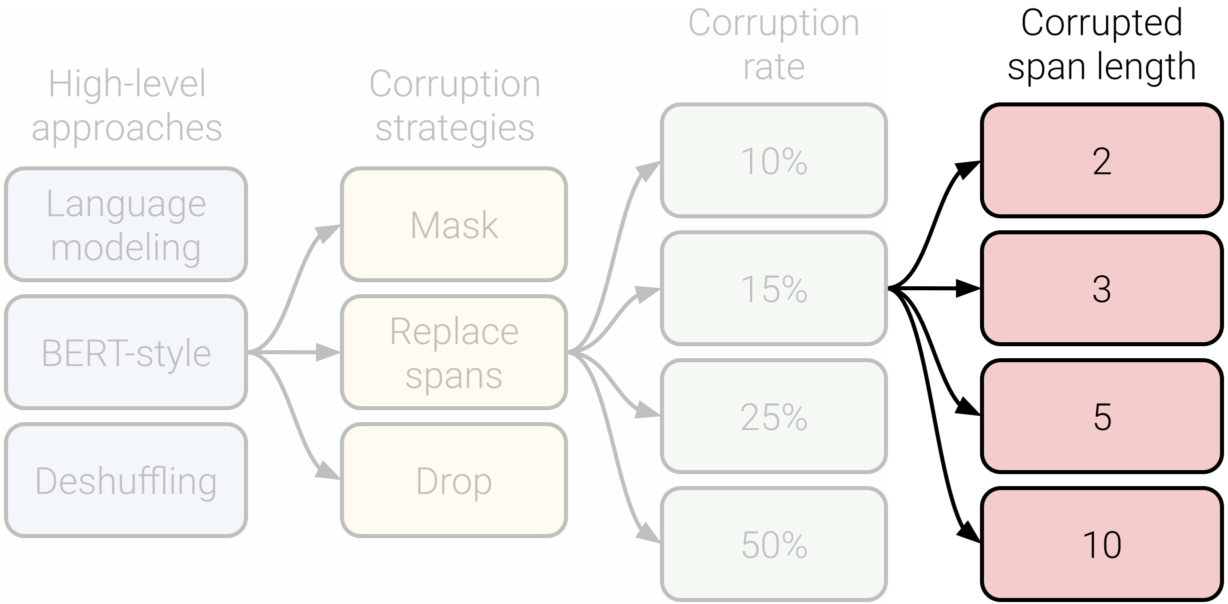
Objective	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
BERT-style [Devlin et al., 2018]	82.96	19.17	**80.65**	69.85	26.78	**40.03**	27.41
MASS-style [Song et al., 2019]	82.32	19.16	80.10	69.28	26.79	**39.89**	27.55
★ Replace corrupted spans	83.28	**19.24**	**80.88**	**71.36**	**26.98**	39.82	**27.65**
Drop corrupted tokens	**84.44**	**19.31**	**80.52**	68.67	**27.07**	39.76	**27.82**



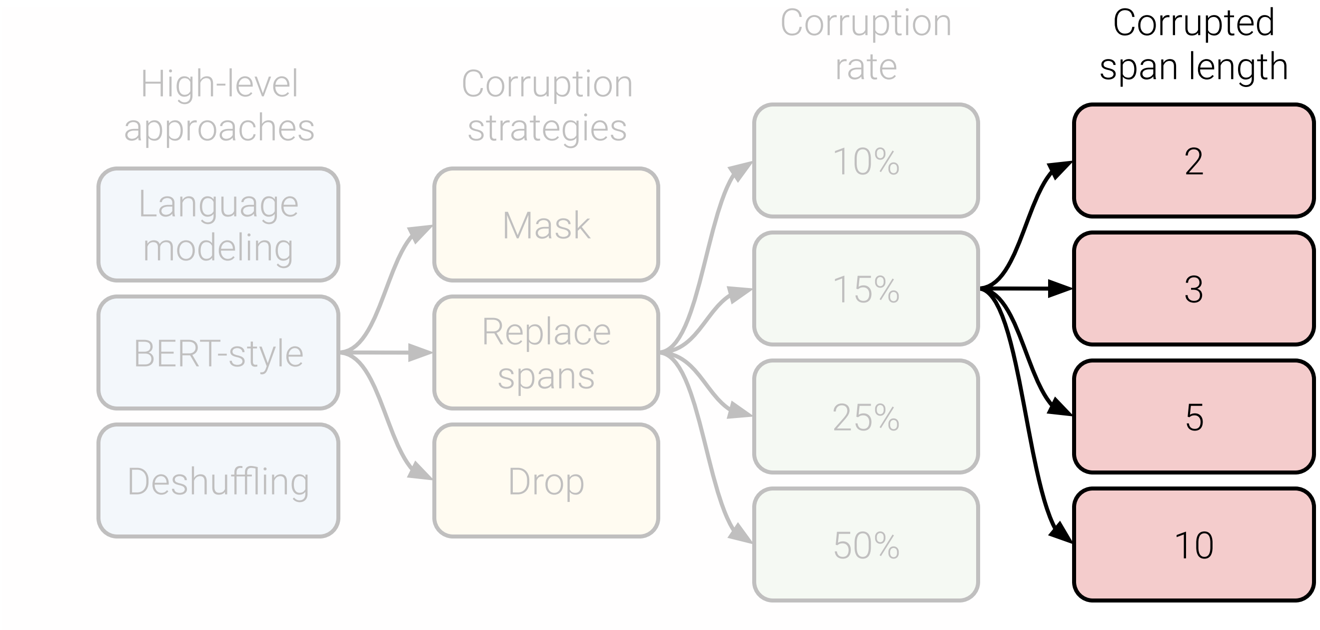
Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
I.i.d. noise, mask tokens	Thank you <M> <M> me to your party <M> week .	*(original text)*
I.i.d. noise, replace spans	**Thank you <X> me to your party <Y> week .**	**<X> for inviting <Y> last <Z>**
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	Thank you <X> to <Y> week .	<X> for inviting me <Y> your party last <Z>



Corruption rate	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
10%	**82.82**	19.00	**80.38**	69.55	**26.87**	39.28	**27.44**
★ 15%	**83.28**	19.24	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
25%	**83.00**	**19.54**	**80.96**	70.48	**27.04**	**39.83**	**27.47**
50%	81.27	19.32	79.80	70.33	**27.01**	**39.90**	**27.49**



Objective	Inputs	Targets
Prefix language modeling	Thank you for inviting	me to your party last week .
BERT-style	Thank you <M> <M> me to your party apple week .	*(original text)*
Deshuffling	party me for your to . last fun you inviting week Thank	*(original text)*
I.i.d. noise, mask tokens	Thank you <M> <M> me to your party <M> week .	*(original text)*
I.i.d. noise, replace spans	Thank you <X> me to your party <Y> week .	<X> for inviting <Y> last <Z>
I.i.d. noise, drop tokens	Thank you me to your party week .	for inviting last
Random spans	**Thank you <X> to <Y> week .**	**<X> for inviting me <Y> your party last <Z>**



Span length	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline (i.i.d.)	**83.28**	19.24	80.88	71.36	**26.98**	**39.82**	**27.65**
2	**83.54**	19.39	**82.09**	**72.20**	**26.76**	**39.99**	**27.63**
3	**83.49**	**19.62**	**81.84**	**72.53**	**26.86**	39.65	**27.62**
5	**83.40**	19.24	**82.05**	**72.23**	**26.88**	39.40	**27.53**
10	82.85	19.33	**81.84**	70.44	**26.79**	39.49	**27.69**

Datasets

Dataset	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**

Menu

Lemon

Introduction

The lemon, Citrus Limon (l.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae.
The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses.
The juice of the lemon is about 5% to 6% citric acid, with a ph of around 2.2, giving it a sour taste.

Please enable JavaScript to use our site.

Home
About
Products
Shipping
Contact
FAQ

Dried Lemons, $3.59/pound

Organic dried lemons from our farm in California.
Lemons are harvested and sun-dried for maximum flavor.
Good in soups and on popcorn.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Curabitur in tempus quam. In mollis et ante at consectetur.
Aliquam erat volutpat.
Duis semper, magna tempor interdum suscipit, ante elit molestie urna, eget efficitur risus nunc ac elit.

```
function Ball(r) {
  this.radius = r;
  this.area = pi * r ** 2;
  this.show = function(){
    drawCircle(r);
  }
}
```

Dataset	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21

Menu

Lemon

Introduction

The lemon, Citrus Limon (l.) Osbeck, is a species of small evergreen tree in the flowering plant family rutaceae.
The tree's ellipsoidal yellow fruit is used for culinary and non-culinary purposes throughout the world, primarily for its juice, which has both culinary and cleaning uses.
The juice of the lemon is about 5% to 6% citric acid, with a ph of around 2.2, giving it a sour taste.

Please enable JavaScript to use our site.

Home
About
Products
Shipping
Contact
FAQ

Dried Lemons, $3.59/pound

Organic dried lemons from our farm in California.
Lemons are harvested and sun-dried for maximum flavor.
Good in soups and on popcorn.

Lorem ipsum dolor sit amet, consectetur adipiscing elit.
Curabitur in tempus quam. In mollis et ante at consectetur.
Aliquam erat volutpat.
Duis semper, magna tempor interdum suscipit, ante elit molestie urna, eget efficitur risus nunc ac elit.

```
function Ball(r) {
  this.radius = r;
  this.area = pi * r ** 2;
  this.show = function(){
    drawCircle(r);
  }
}
```

Dataset	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**

48-year old Alain Robert, affectionately known as the 'French Spiderman', has climbed a 57-storey Sydney skyscraper without any equipment in 20 minutes.
The purpose of Alain Robert's actions was to raise awareness of global warming.
Following the previous like events in other cities, he was arrested and will possibly be fined.
When Robert was 12, he climbed eight storeys to get into his flat instead of waiting for his parents to return.
Since then, he has climbed over eighty buildings around the world, including the Eiffel Tower, The New York Times building, and Sydney Harbour Bridge.

The Conservative Party has won the last seat of the 2010 general election, taking the seat of Thirsk and Malton with a majority of over 11,000.
Voting in the constituency had been delayed by the death of the United Kingdom Independence Party candidate in the run-up to the original polling date of May 6, 2010.
The new MP, Anne McIntosh, took over 52% of the vote, with the Liberal Democrats – partners with the Conservatives in the coalition government – in second place.
Despite the relationship between the parties in government, the Liberal Democrat candidate Howard Keal had promised that there would be a "full-on fight" for the seat.

Melbourne writer Harry Nicolaides, 41, was sentenced on Monday to three years imprisonment for defaming the Royal Family of Thailand.
He had pled guilty to the lèse majesté indictment that arose from a self-published 2005 novel, Verisimilitude, of which only 50 copies were printed, and just seven sold.
Meanwhile, yesterday, the Thai police charged a leading leftist political science professor, Dr. Giles Ji Ungpakorn, with lèse majesté.
The passage of concern, which comprised only 103 words or 12 lines, referred to a crown prince's love life.

	Dataset	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★	C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
	C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
	RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
	WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**

48-year old Alain Robert, affectionately known as the 'French Spiderman', has climbed a 57-storey Sydney skyscraper without any equipment in 20 minutes.
The purpose of Alain Robert's actions was to raise awareness of global warming.
Following the previous like events in other cities, he was arrested and will possibly be fined.
When Robert was 12, he climbed eight storeys to get into his flat instead of waiting for his parents to return.
Since then, he has climbed over eighty buildings around the world, including the Eiffel Tower, The New York Times building, and Sydney Harbour Bridge.

Sheila Mullen is a Scottish painter who lives and works in Scotland.
She was born on 24 January 1942 in Glasgow, Scotland.
She grew up near Auchtermuchty, Fife, Scotland.
She attended the Glasgow School of Art and started painting professionally in 1978.
Her works are in the permanent collections of the Leeds Art Gallery and the Duke of Buccleuch among others. In 2010 she was the subject of a monograph by Ann Matheson: The Bairns O Adam: The Paintings of Sheila Mullen.
In 2006 she collaborated with the group of Scottish writers called the Crichton Writers.

Purchasing distressed jeans from a store can get pricey.
Luckily, you can turn a regular pair of jeans into distressed jeans with a shaving razor or pair of scissors.
It's fun and easy to turn even the oldest, most unfashionable pair of jeans into distressed jeans.
Mark the areas you want to distress.
Then, cut horizontal lines with the razor and remove threads with a pair of tweezers.
When you're done, you'll have your own pair of distressed jeans.
1. Select the jeans you want to distress.
Choose a pair of jeans you do not mind ripping or fraying.

Dataset	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**
Wikipedia	16GB	81.85	**19.31**	81.29	68.01	**26.94**	39.69	**27.67**

The Templo Expiatorio del Santísimo Sacramento is a Catholic church dedicated to the Blessed Sacrament, located in Guadalajara, Jalisco, Mexico.
The church is considered the greatest work of its kind in Mexico.
Its construction began on August 15, 1897 and ended 75 years later in 1972.
The idea of building a temple dedicated to the Blessed Sacrament in the city of Guadalajara began in the late 19th century, when a congregation of Catholics formed a committee to make this temple.
The construction began with a ceremony hosted by Archbishop Pedro Loza y Pardavé and Pedro Romero.

Sheila Mullen is a Scottish painter who lives and works in Scotland.
She was born on 24 January 1942 in Glasgow, Scotland.
She grew up near Auchtermuchty, Fife, Scotland.
She attended the Glasgow School of Art and started painting professionally in 1978.
Her works are in the permanent collections of the Leeds Art Gallery and the Duke of Buccleuch among others. In 2010 she was the subject of a monograph by Ann Matheson: The Bairns O Adam: The Paintings of Sheila Mullen.
In 2006 she collaborated with the group of Scottish writers called the Crichton Writers.

The DB Museum in Koblenz was opened on 21 April 2001 as the first remote site of the Nuremberg Transport Museum.
It is run by volunteer workers as part of the Stiftung Bahn-Sozialwerk (BSW), a kind of railway workers social service organisation, and has its origins in a BSW's 'Group for the Preservation of Historical Railway Vehicles' at Koblenz.
The DB Museum, Koblenz, is housed in the former goods wagon repair shop (Ausbesserungswerk) in the Koblenz district of Lützel. The site was built in 1905 as part of the rebuilding and expansion of Lützel goods station into the Koblenz-Lützel locomotive depot (Bahnbetriebswerk).

Dataset	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**
Wikipedia	16GB	81.85	**19.31**	81.29	68.01	**26.94**	39.69	**27.67**
Wikipedia + TBC	20GB	83.65	**19.28**	**82.08**	**73.24**	**26.77**	39.63	**27.57**

The Templo Expiatorio del Santísimo Sacramento is a Catholic church dedicated to the Blessed Sacrament, located in Guadalajara, Jalisco, Mexico.
The church is considered the greatest work of its kind in Mexico.
Its construction began on August 15, 1897 and ended 75 years later in 1972.
The idea of building a temple dedicated to the Blessed Sacrament in the city of Guadalajara began in the late 19th century, when a congregation of Catholics formed a committee to make this temple.
The construction began with a ceremony hosted by Archbishop Pedro Loza y Pardavé and Pedro Romero.

Sheila Mullen is a Scottish painter who lives and works in Scotland.
She was born on 24 January 1942 in Glasgow, Scotland.
She grew up near Auchtermuchty, Fife, Scotland.
She attended the Glasgow School of Art and started painting professionally in 1978.
Her works are in the permanent collections of the Leeds Art Gallery and the Duke of Buccleuch among others. In 2010 she was the subject of a monograph by Ann Matheson: The Bairns O Adam: The Paintings of Sheila Mullen.
In 2006 she collaborated with the group of Scottish writers called the Crichton Writers.

Down they went, feeling a trifle timid, for they seldom went to parties, and informal as this little gathering was, it was an event to them.
Mrs. Gardiner, a stately old lady, greeted them kindly and handed them over to the eldest of her six daughters.
Meg knew Sallie and was at her ease very soon, but Jo, who didn't care much for girls or girlish gossip, stood about, with her back carefully against the wall, and felt as much out of place as a colt in a flower garden.
Half a dozen jovial lads were talking about skates in another part of the room, and she longed to go and join them, for skating was one of the joys of her life.

Dataset	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**
Wikipedia	16GB	81.85	**19.31**	81.29	68.01	**26.94**	39.69	**27.67**
Wikipedia + TBC	20GB	83.65	**19.28**	**82.08**	**73.24**	**26.77**	39.63	**27.57**

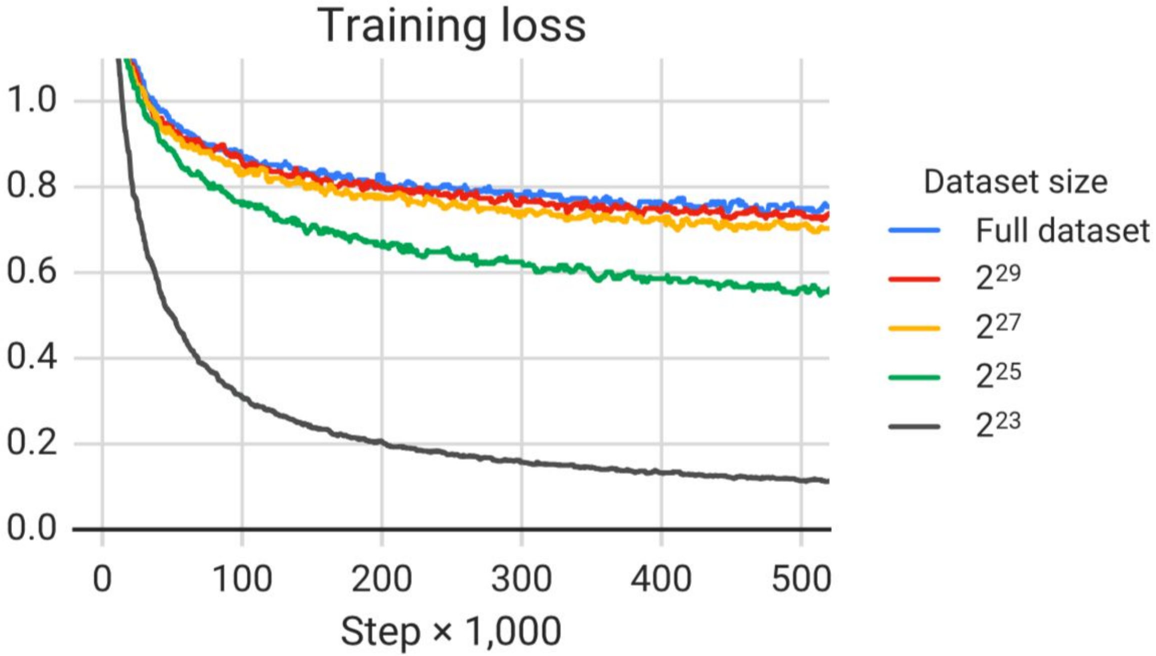
Much worse on CoLA

Much better on ReCoRD

Order of magnitude smaller

Much better on MultiRC

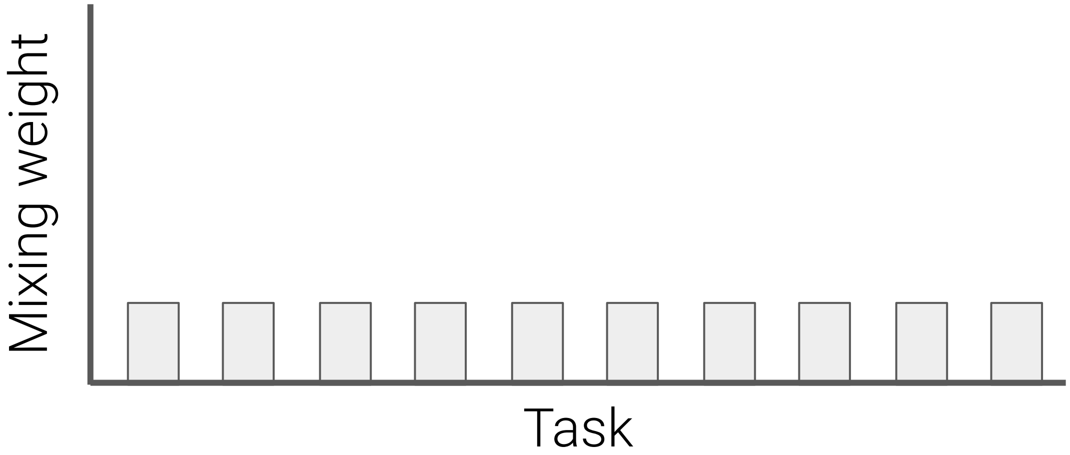
Number of tokens	Repeats	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Full dataset	0	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
2^{29}	64	**82.87**	**19.19**	**80.97**	**72.03**	**26.83**	**39.74**	**27.63**
2^{27}	256	82.62	**19.20**	79.78	69.97	**27.02**	**39.71**	27.33
2^{25}	1,024	79.55	18.57	76.27	64.76	26.38	39.56	26.80
2^{23}	4,096	76.34	18.33	70.92	59.29	26.37	38.84	25.81



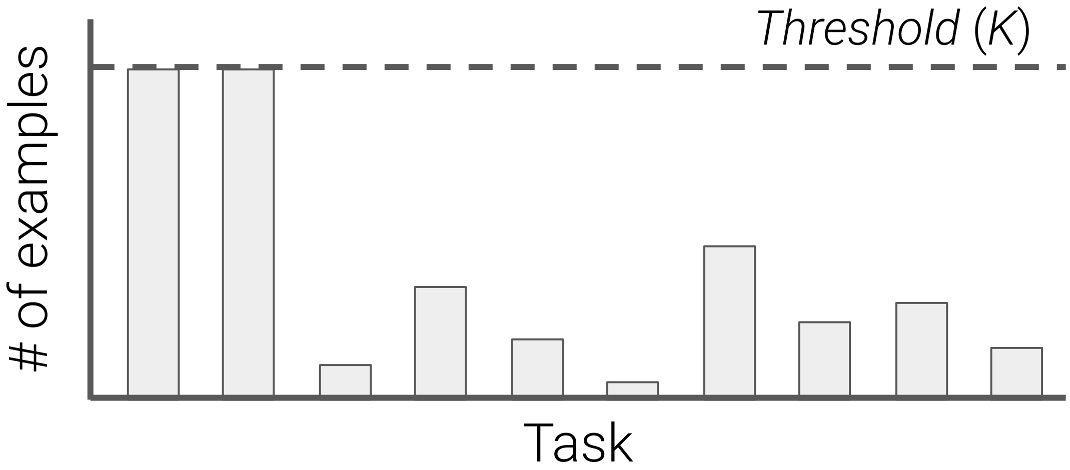
Multi-task

Mixing strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline (pre-train/fine-tine)	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**

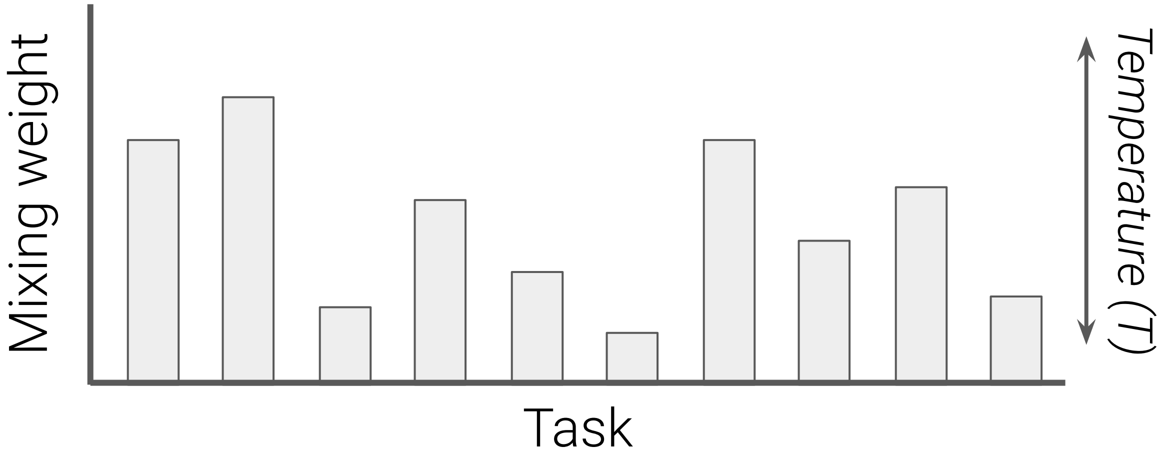
Mixing strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline (pre-train/fine-tine)	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Equal	76.13	19.02	76.51	63.37	23.89	34.31	26.78



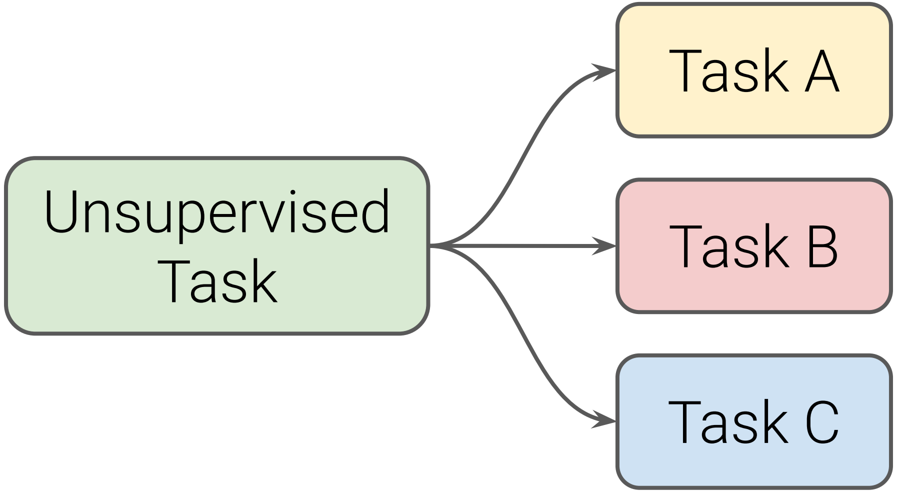
Mixing strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline (pre-train/fine-tine)	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Equal	76.13	19.02	76.51	63.37	23.89	34.31	26.78
Examples-proportional, $K = 2^{16}$	80.45	19.04	77.25	69.95	24.35	34.99	27.10
Examples-proportional, $K = 2^{17}$	81.56	19.12	77.00	67.91	24.36	35.00	27.25
Examples-proportional, $K = 2^{18}$	81.67	19.07	78.17	67.94	24.57	35.19	27.39
Examples-proportional, $K = 2^{19}$	81.42	**19.24**	79.78	67.30	25.21	36.30	**27.76**
Examples-proportional, $K = 2^{20}$	80.80	**19.24**	**80.36**	67.38	25.66	36.93	**27.68**
Examples-proportional, $K = 2^{21}$	79.83	18.79	79.50	65.10	25.82	37.22	27.13



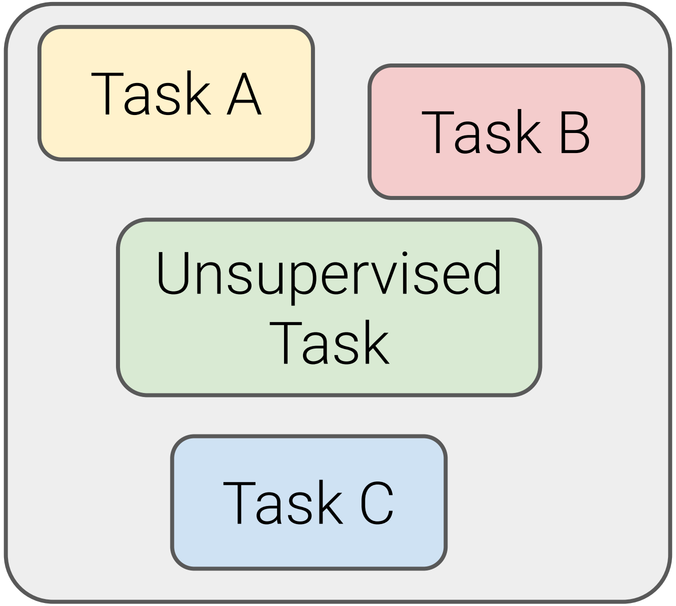
Mixing strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline (pre-train/fine-tine)	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Equal	76.13	19.02	76.51	63.37	23.89	34.31	26.78
Examples-proportional, $K = 2^{16}$	80.45	19.04	77.25	69.95	24.35	34.99	27.10
Examples-proportional, $K = 2^{17}$	81.56	19.12	77.00	67.91	24.36	35.00	27.25
Examples-proportional, $K = 2^{18}$	81.67	19.07	78.17	67.94	24.57	35.19	27.39
Examples-proportional, $K = 2^{19}$	81.42	**19.24**	79.78	67.30	25.21	36.30	**27.76**
Examples-proportional, $K = 2^{20}$	80.80	**19.24**	**80.36**	67.38	25.66	36.93	**27.68**
Examples-proportional, $K = 2^{21}$	79.83	18.79	79.50	65.10	25.82	37.22	27.13
Temperature-scaled, $T = 2$	81.90	**19.28**	79.42	69.92	25.42	36.72	27.20
Temperature-scaled, $T = 4$	80.56	**19.22**	77.99	69.54	25.04	35.82	27.45
Temperature-scaled, $T = 8$	77.21	19.10	77.14	66.07	24.55	35.35	27.17



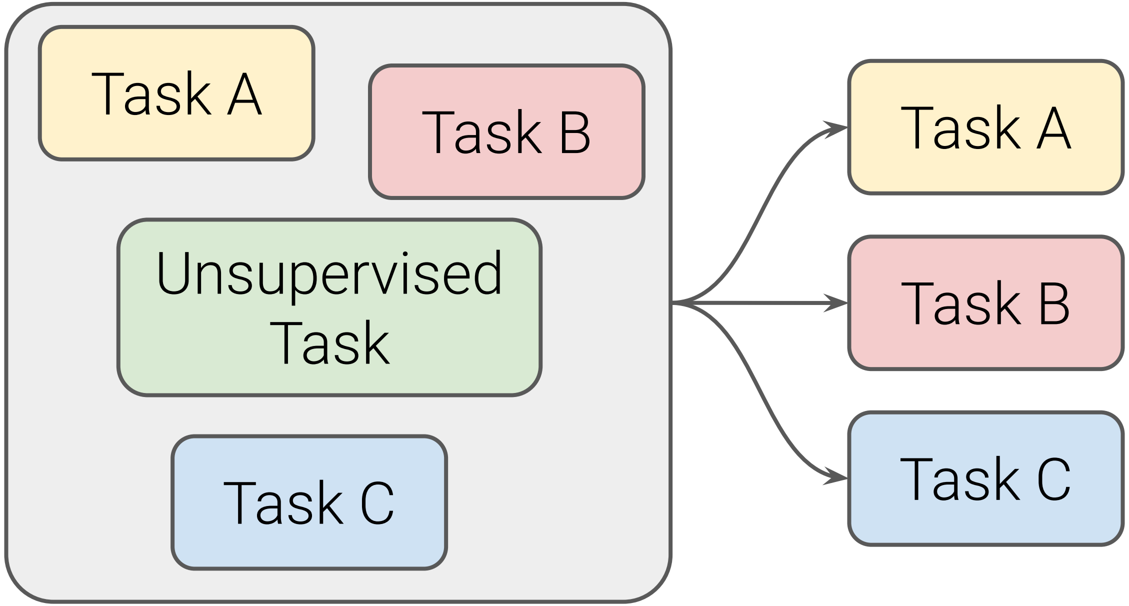
Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65



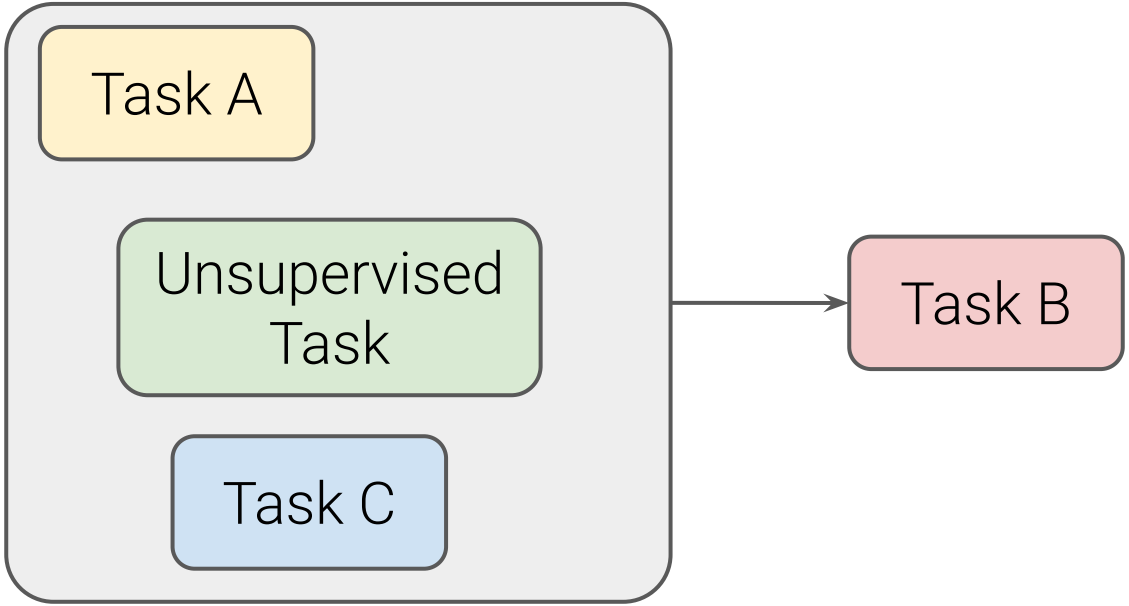
Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76



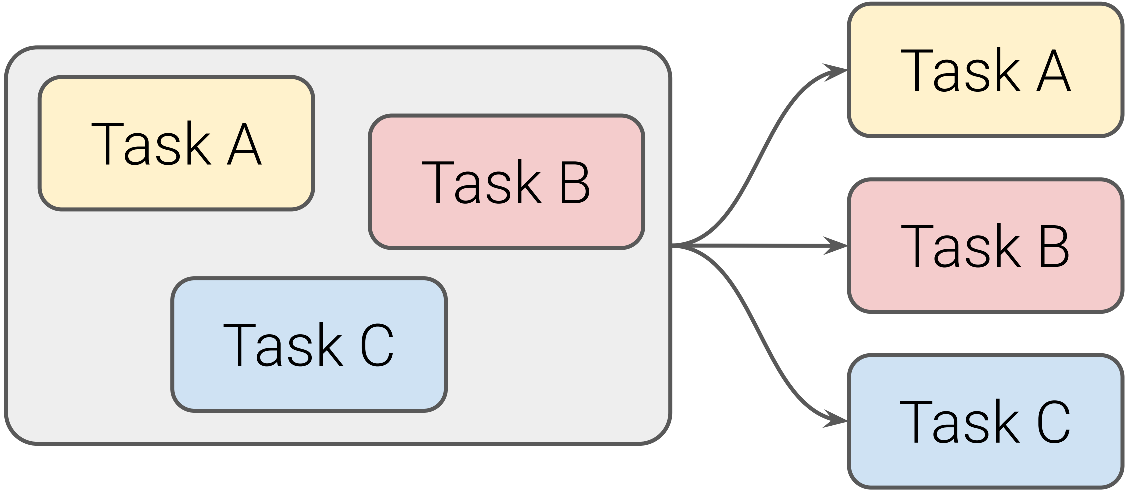
Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76
Multi-task pre-training + fine-tuning	**83.11**	**19.12**	**80.26**	**71.03**	**27.08**	39.80	**28.07**



Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76
Multi-task pre-training + fine-tuning	**83.11**	**19.12**	**80.26**	**71.03**	**27.08**	39.80	**28.07**
Leave-one-out multi-task training	81.98	19.05	79.97	**71.68**	**26.93**	39.79	**27.87**



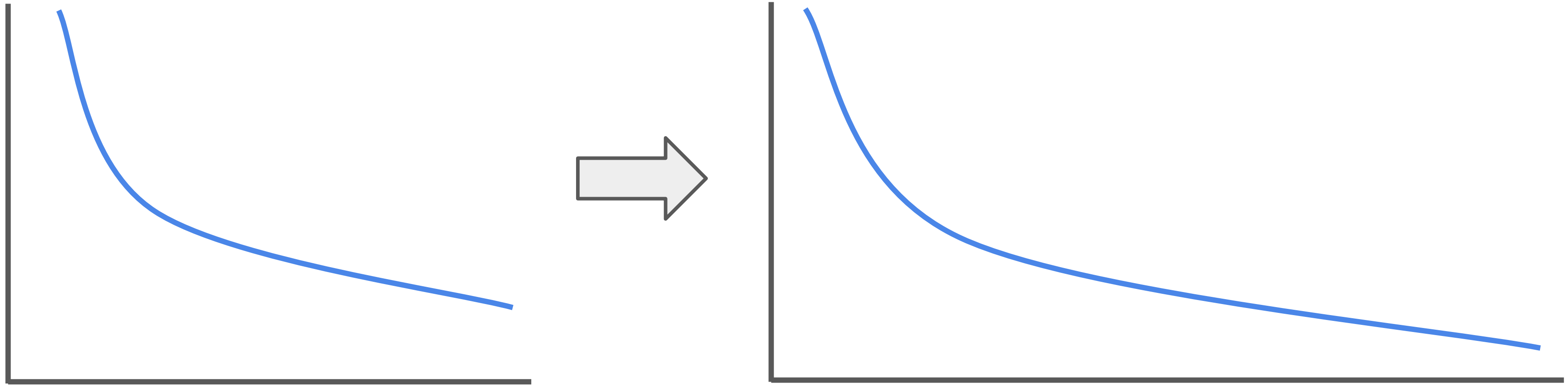
Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76
Multi-task pre-training + fine-tuning	**83.11**	**19.12**	**80.26**	**71.03**	**27.08**	39.80	**28.07**
Leave-one-out multi-task training	81.98	19.05	79.97	**71.68**	**26.93**	39.79	**27.87**
Supervised multi-task pre-training	79.93	18.96	77.38	65.36	26.81	**40.13**	**28.04**



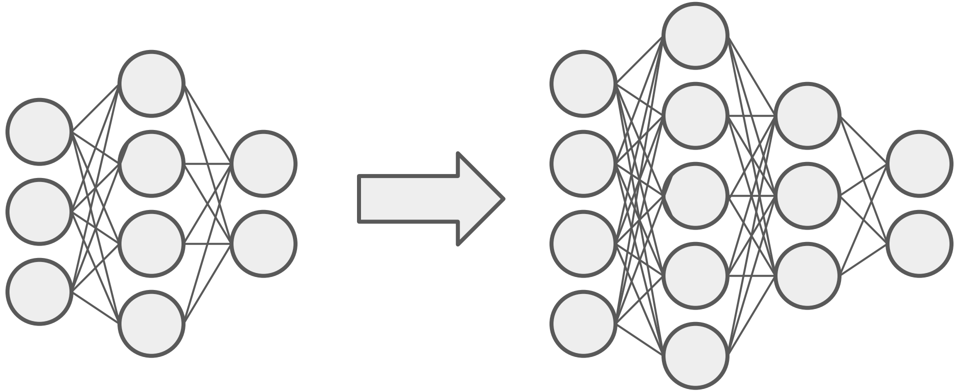
Scaling

Scaling strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline	83.28	19.24	80.88	71.36	26.98	39.82	27.65

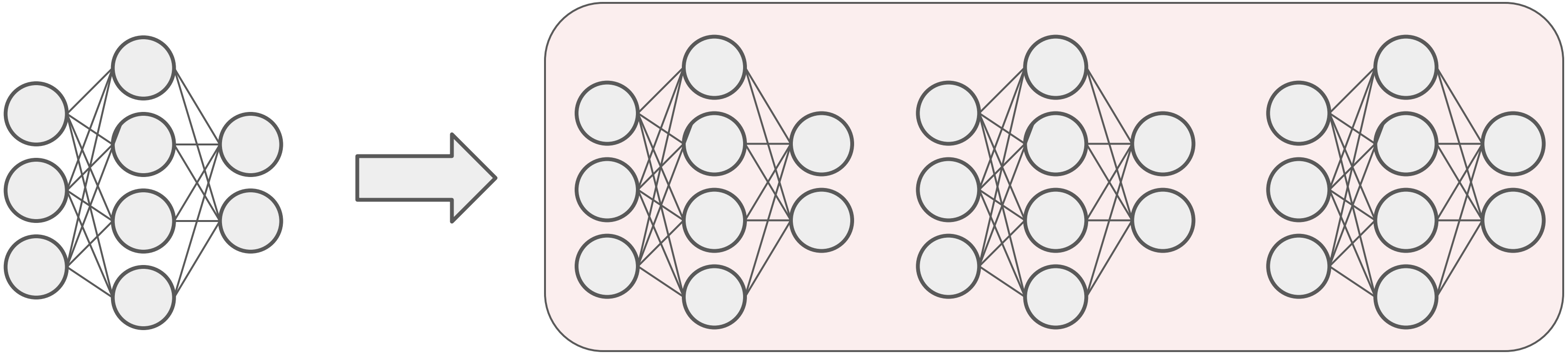
Scaling strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline	83.28	19.24	80.88	71.36	26.98	39.82	27.65
1× size, 4× training steps	85.33	19.33	82.45	74.72	27.08	40.66	27.93
1× size, 4× batch size	84.60	19.42	82.52	74.64	27.07	40.60	27.84



Scaling strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline	83.28	19.24	80.88	71.36	26.98	39.82	27.65
1× size, 4× training steps	85.33	19.33	82.45	74.72	27.08	40.66	27.93
1× size, 4× batch size	84.60	19.42	82.52	74.64	27.07	40.60	27.84
2× size, 2× training steps	**86.18**	19.66	**84.18**	77.18	27.52	**41.03**	28.19
4× size, 1× training steps	**85.91**	19.73	**83.86**	**78.04**	27.47	40.71	28.10



Scaling strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline	83.28	19.24	80.88	71.36	26.98	39.82	27.65
1× size, 4× training steps	85.33	19.33	82.45	74.72	27.08	40.66	27.93
1× size, 4× batch size	84.60	19.42	82.52	74.64	27.07	40.60	27.84
2× size, 2× training steps	**86.18**	19.66	**84.18**	77.18	27.52	**41.03**	28.19
4× size, 1× training steps	**85.91**	19.73	**83.86**	**78.04**	27.47	40.71	28.10
4× ensembled	84.77	**20.10**	83.09	71.74	**28.05**	40.53	**28.57**
4× ensembled, fine-tune only	84.05	19.57	82.36	71.55	27.55	40.22	28.09



Putting it all together

Encoder-decoder architecture

Architecture	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	$2P$	M	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Enc-dec, shared	P	M	82.81	18.78	**80.63**	**70.73**	26.72	39.03	**27.46**
Enc-dec, 6 layers	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39

Span prediction objective

Span length	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline (i.i.d.)	**83.28**	19.24	80.88	71.36	**26.98**	**39.82**	**27.65**
2	**83.54**	19.39	**82.09**	**72.20**	**26.76**	**39.99**	**27.63**
3	**83.49**	**19.62**	**81.84**	**72.53**	**26.86**	39.65	**27.62**
5	**83.40**	19.24	**82.05**	**72.23**	**26.88**	39.40	**27.53**
10	82.85	19.33	**81.84**	70.44	**26.79**	39.49	**27.69**

C4 dataset

Dataset	Size	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ C4	745GB	83.28	**19.24**	80.88	71.36	**26.98**	**39.82**	**27.65**
C4, unfiltered	6.1TB	81.46	19.14	78.78	68.04	26.55	39.34	27.21
RealNews-like	35GB	**83.83**	**19.23**	80.39	72.38	**26.75**	**39.90**	**27.48**
WebText-like	17GB	**84.03**	**19.31**	**81.42**	71.40	**26.80**	**39.74**	**27.59**
Wikipedia	16GB	81.85	**19.31**	81.29	68.01	**26.94**	39.69	**27.67**
Wikipedia + TBC	20GB	83.65	**19.28**	**82.08**	**73.24**	**26.77**	39.63	**27.57**

Multi-task pre-training

Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76
Multi-task pre-training + fine-tuning	**83.11**	**19.12**	**80.26**	**71.03**	**27.08**	39.80	**28.07**
Leave-one-out multi-task training	81.98	19.05	79.97	**71.68**	**26.93**	39.79	**27.87**
Supervised multi-task pre-training	79.93	18.96	77.38	65.36	26.81	**40.13**	**28.04**

Bigger models trained longer

Scaling strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
Baseline	83.28	19.24	80.88	71.36	26.98	39.82	27.65
1× size, 4× training steps	85.33	19.33	82.45	74.72	27.08	40.66	27.93
1× size, 4× batch size	84.60	19.42	82.52	74.64	27.07	40.60	27.84
2× size, 2× training steps	**86.18**	19.66	**84.18**	77.18	27.52	**41.03**	28.19
4× size, 1× training steps	**85.91**	19.73	**83.86**	**78.04**	27.47	40.71	28.10
4× ensembled	84.77	**20.10**	83.09	71.74	**28.05**	40.53	**28.57**
4× ensembled, fine-tune only	84.05	19.57	82.36	71.55	27.55	40.22	28.09

Model size variants

Model	Parameters	# layers	d_{model}	d_{ff}	d_{kv}	# heads
Small	60M	6	512	2048	64	8
Base	220M	12	768	3072	64	12
Large	770M	24	1024	4096	64	16
3B	3B	24	1024	16384	128	32
11B	11B	24	1024	65536	128	128

Model	GLUE Average	CNN/DM ROUGE-2-F	SQuAD EM	SuperGLUE Average	WMT EnDe BLEU	WMT EnFr BLEU	WMT EnRo BLEU
Previous best	89.4	20.30	90.1	84.6	**33.8**	**43.8**	**38.5**
T5-Small	77.4	19.56	87.24	63.3	26.7	36.0	26.8
T5-Base	82.7	20.34	92.08	76.2	30.9	41.2	28.0
T5-Large	86.4	20.68	93.79	82.3	32.0	41.5	28.1
T5-3B	88.5	21.02	94.95	86.4	31.8	42.6	28.2
T5-11B	**90.3**	**21.55**	**91.26**	**89.3**	32.1	43.4	28.1

Back-translation beats English-only pre-training

Model	GLUE Average	CNN/DM ROUGE-2-F	SQuAD EM	SuperGLUE Average	WMT EnDe BLEU	WMT EnFr BLEU	WMT EnRo BLEU
Previous best	89.4	20.30	90.1	84.6	**33.8**	**43.8**	**38.5**
T5-Small	77.4	19.56	87.24	63.3	26.7	36.0	26.8
T5-Base	82.7	20.34	92.08	76.2	30.9	41.2	28.0
T5-Large	86.4	20.68	93.79	82.3	32.0	41.5	28.1
T5-3B	88.5	21.02	94.95	86.4	31.8	42.6	28.2
T5-11B	**90.3**	**21.55**	**91.26**	**89.3**	32.1	43.4	28.1

Human score = 89.8

google-research / text-to-text-transfer-transformer

<> Code | Issues 8 | Pull requests 1 | Actions | Projects 0 | Wiki | Security | Insights | Settings

Code for the paper "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" https://arxiv.org/abs/1910.10683

Edit

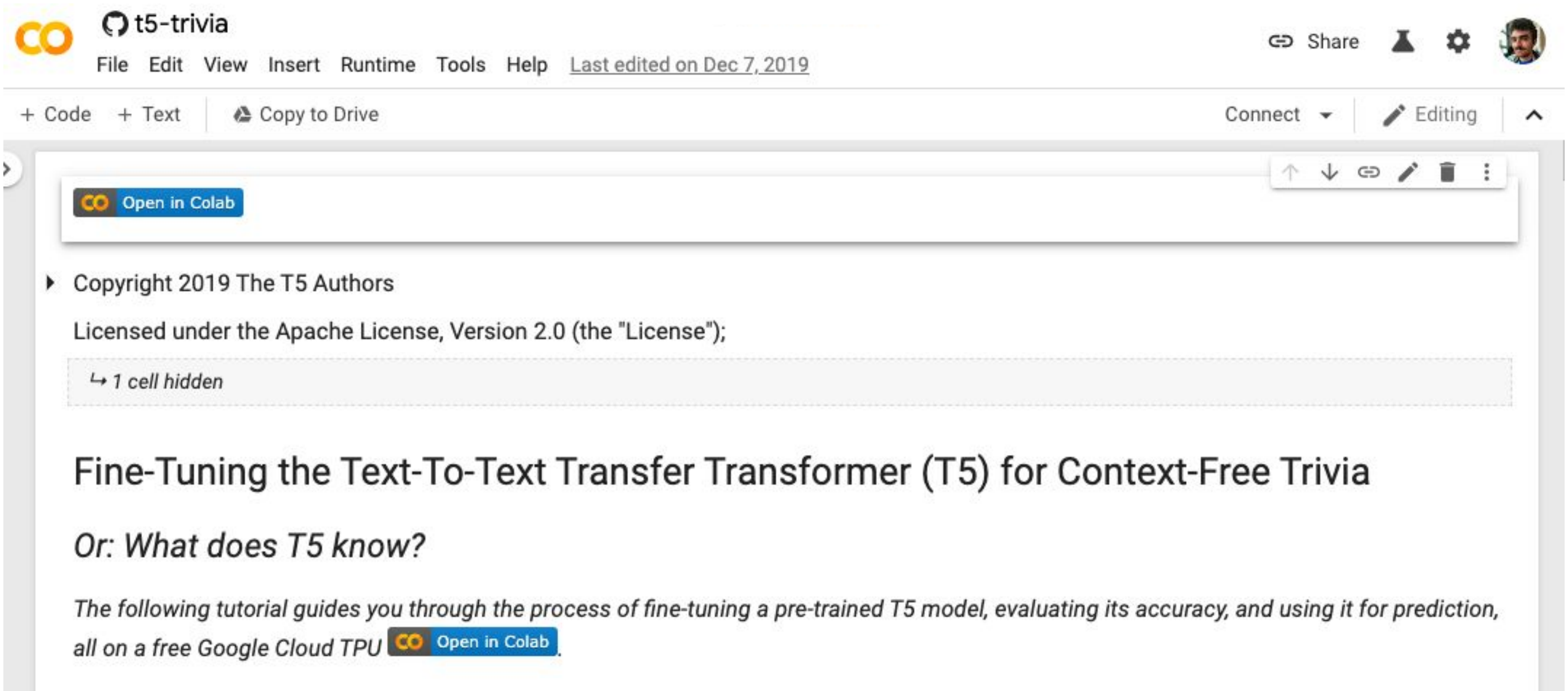
Manage topics

Released Model Checkpoints

We have released the following checkpoints for pre-trained models described in our paper:

- **T5-Small** (60 million parameters): gs://t5-data/pretrained_models/small
- **T5-Base** (220 million parameters): gs://t5-data/pretrained_models/base
- **T5-Large** (770 million parameters): gs://t5-data/pretrained_models/large
- **T5-3B** (3 billion parameters): gs://t5-data/pretrained_models/3B
- **T5-11B** (11 billion parameters): gs://t5-data/pretrained_models/11B

https://github.com/google-research/text-to-text-transfer-transformer



http://tiny.cc/t5-colab

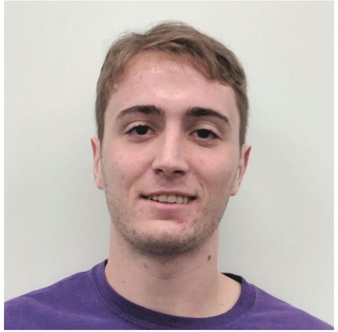

Colin Raffel


Noam Shazeer


Adam Roberts


Katherine Lee

Sharan Narang


Michael Matena


Yanqi Zhou


Wei Li


Peter J. Liu

Questions?