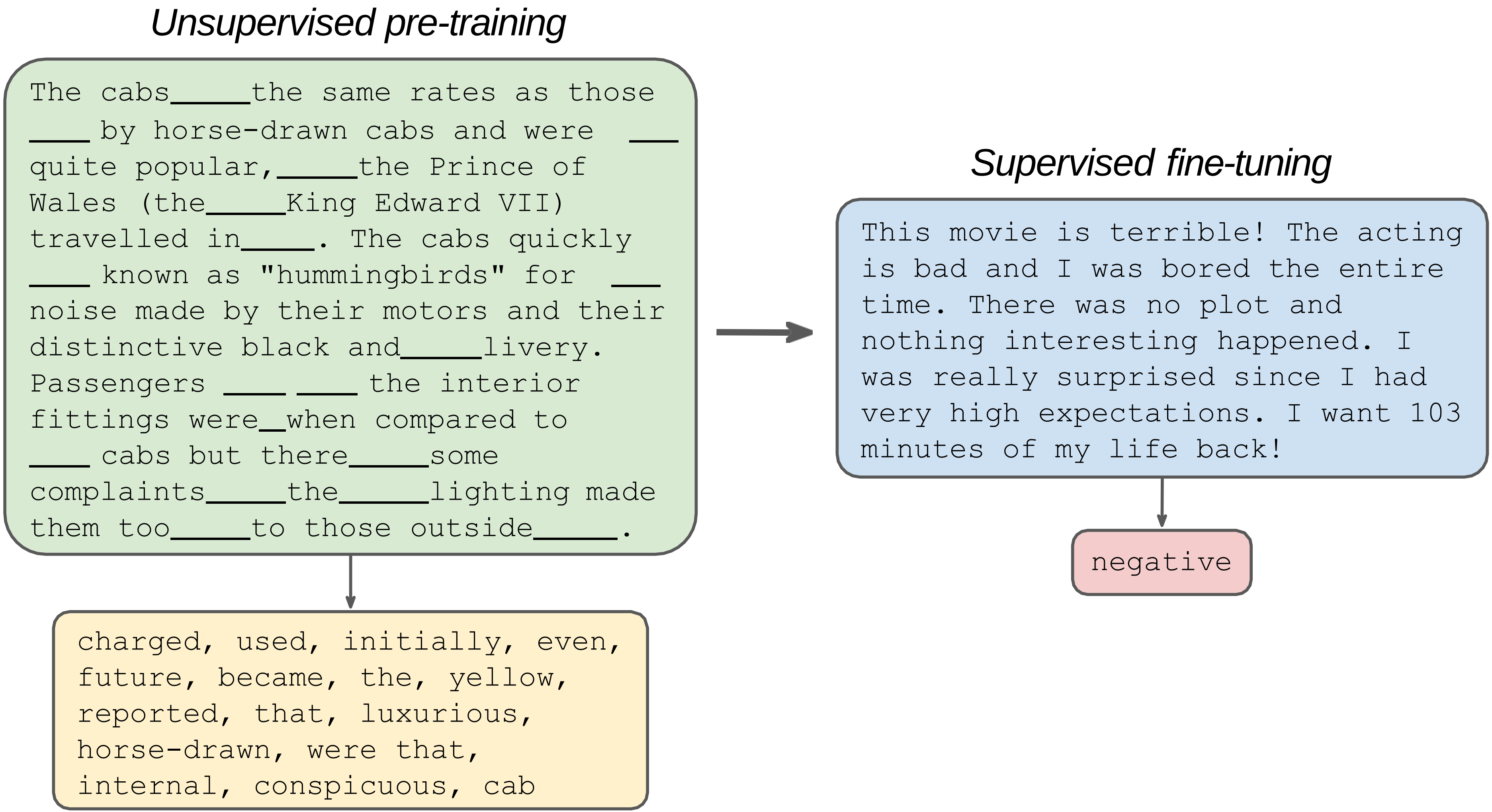
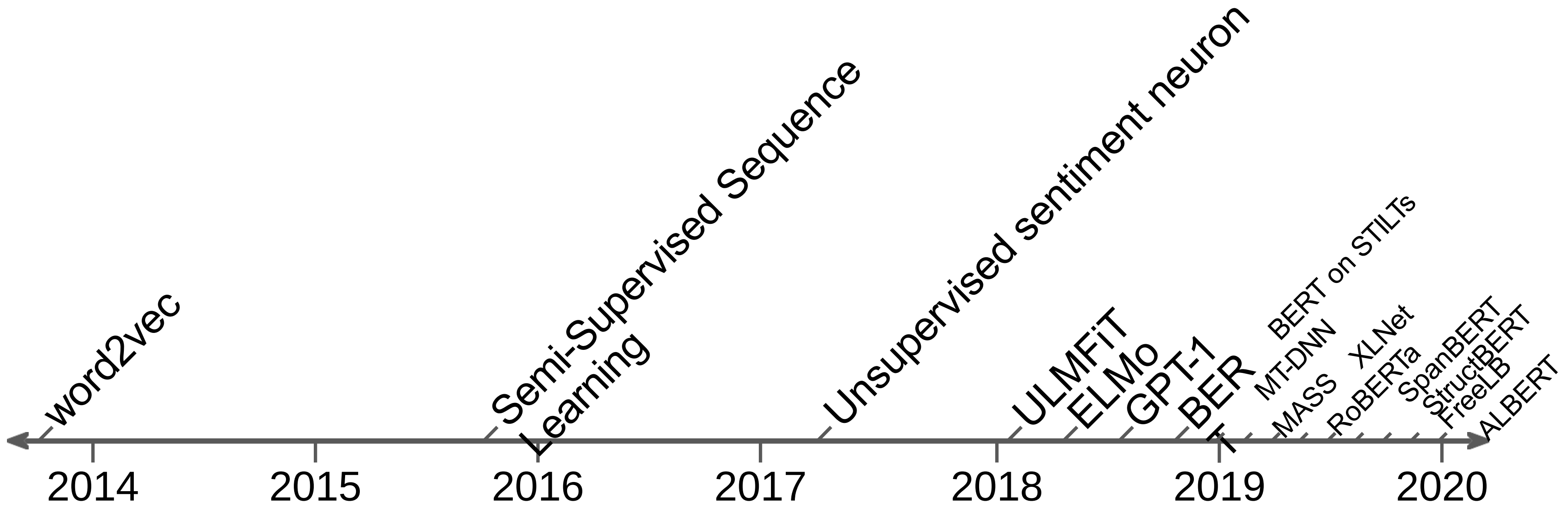
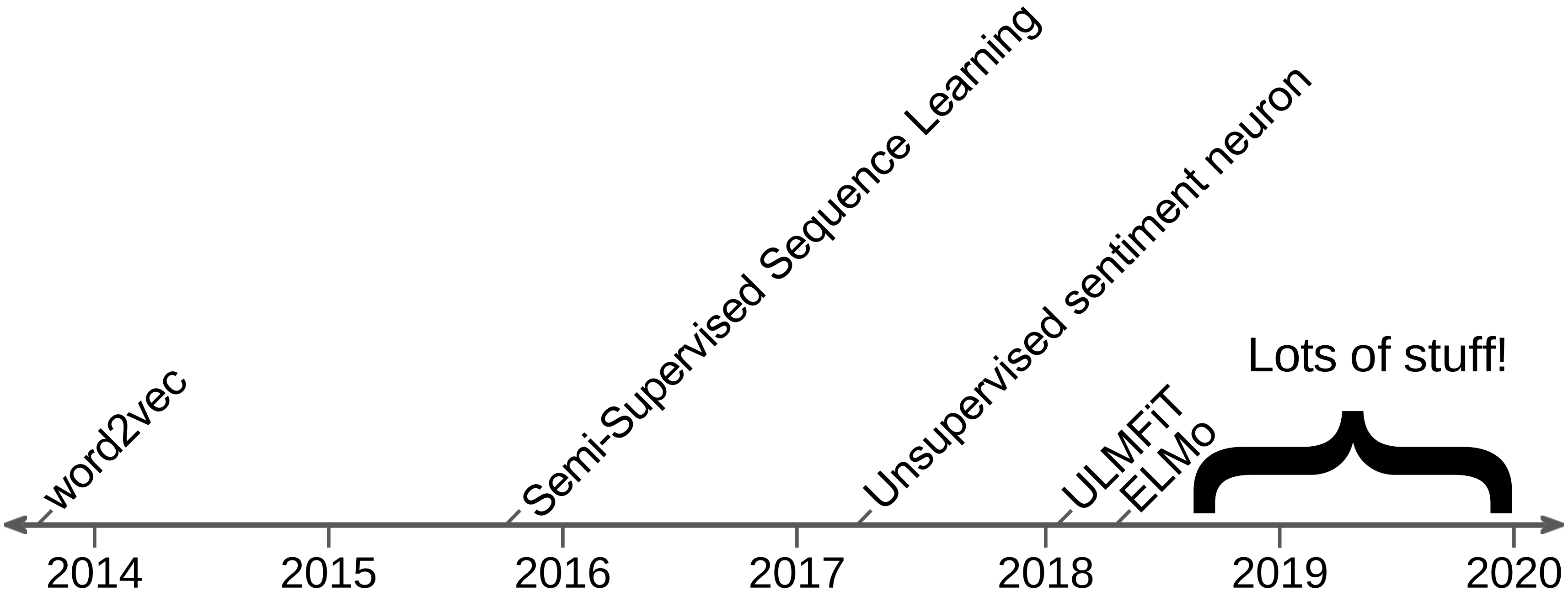
Text-to-Text Transfer and Decoding

Slides are adopted from Colin Raffel, C Manning


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


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

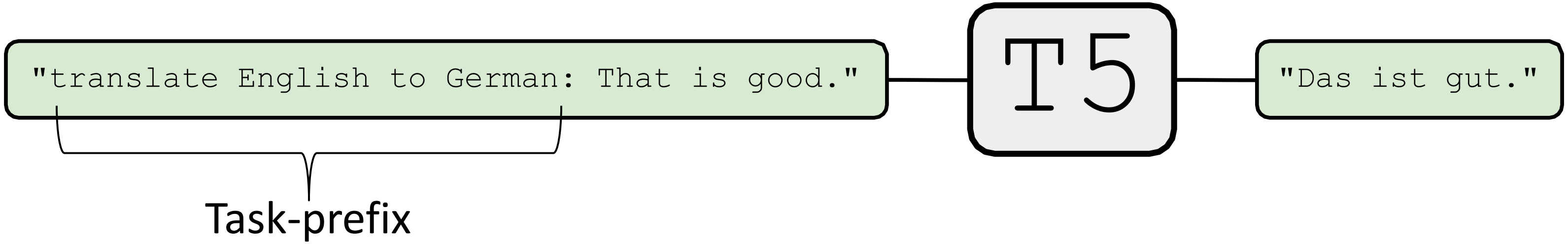
T5

Treating all text problems in the same format

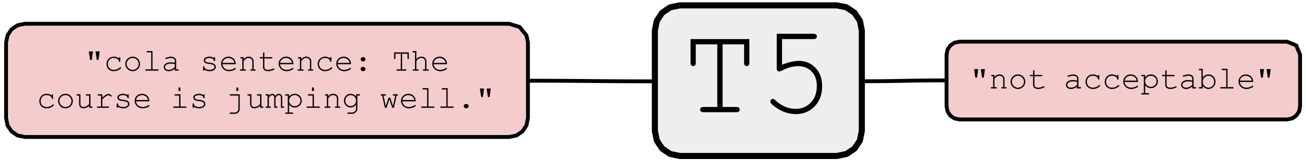
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

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

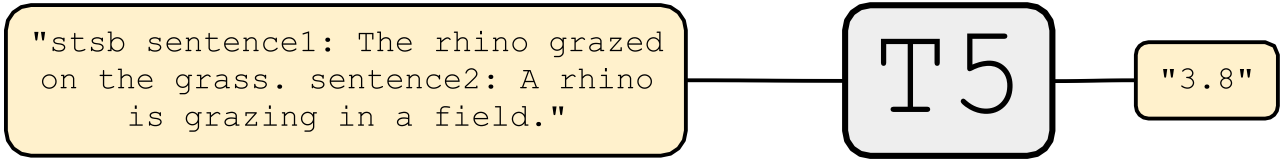
Transfer learning, where a model is first pre-trained on a data-rich task before being fine-tuned on a downstream task, has emerged as a powerful technique in natural language processing (NLP). The effectiveness of transfer learning has given rise to a diversity of approaches, methodology, and practice. In this paper, we explore the landscape of transfer learning techniques for NLP by introducing a unified framework that converts all text-based language problems into a text-to-text format. Our systematic study compares pre-training objectives, architectures, unlabeled data sets, transfer approaches, and other factors on dozens of language understanding tasks. By combining the insights from our exploration with scale and our new ``Colossal Clean Crawled Corpus", we achieve state-of-the-art results on many benchmarks covering summarization, question answering, text classification, and more. To facilitate future work on transfer learning for NLP, we release our data set, pre-trained models, and code.



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

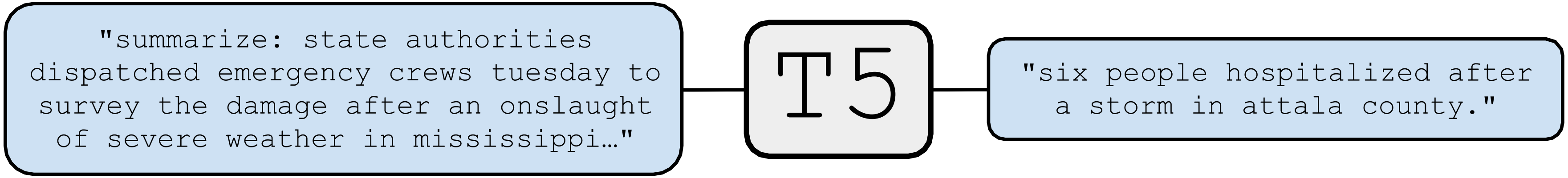


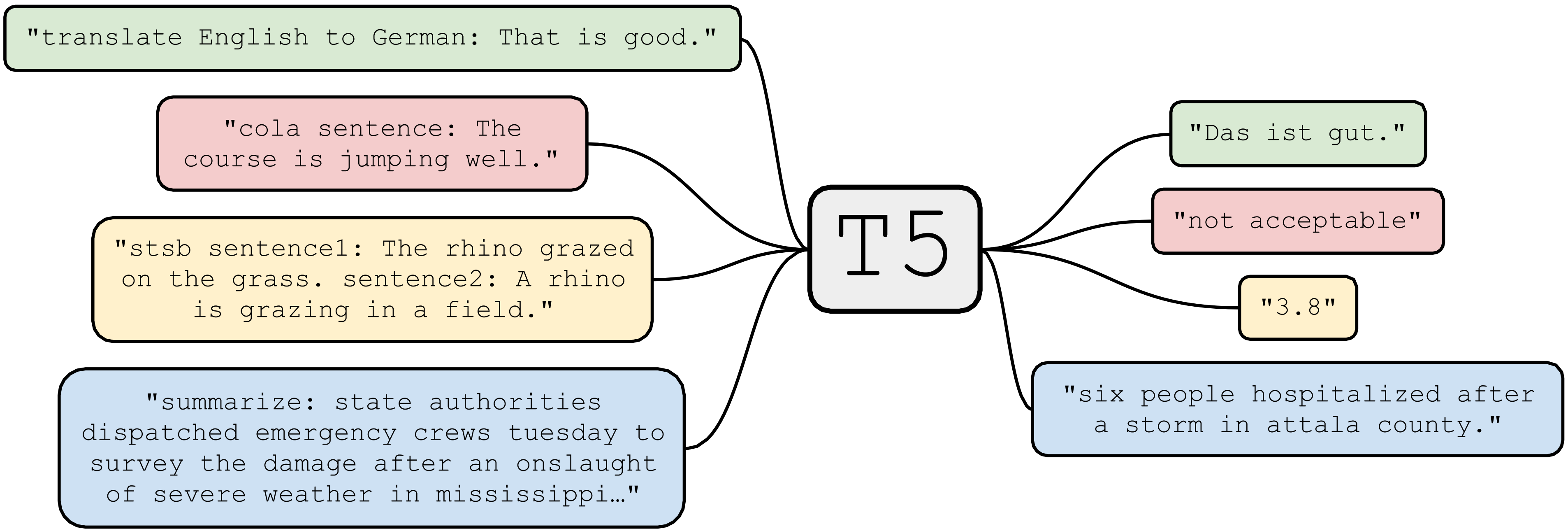
COLA: Corpus of Linguistic Acceptability

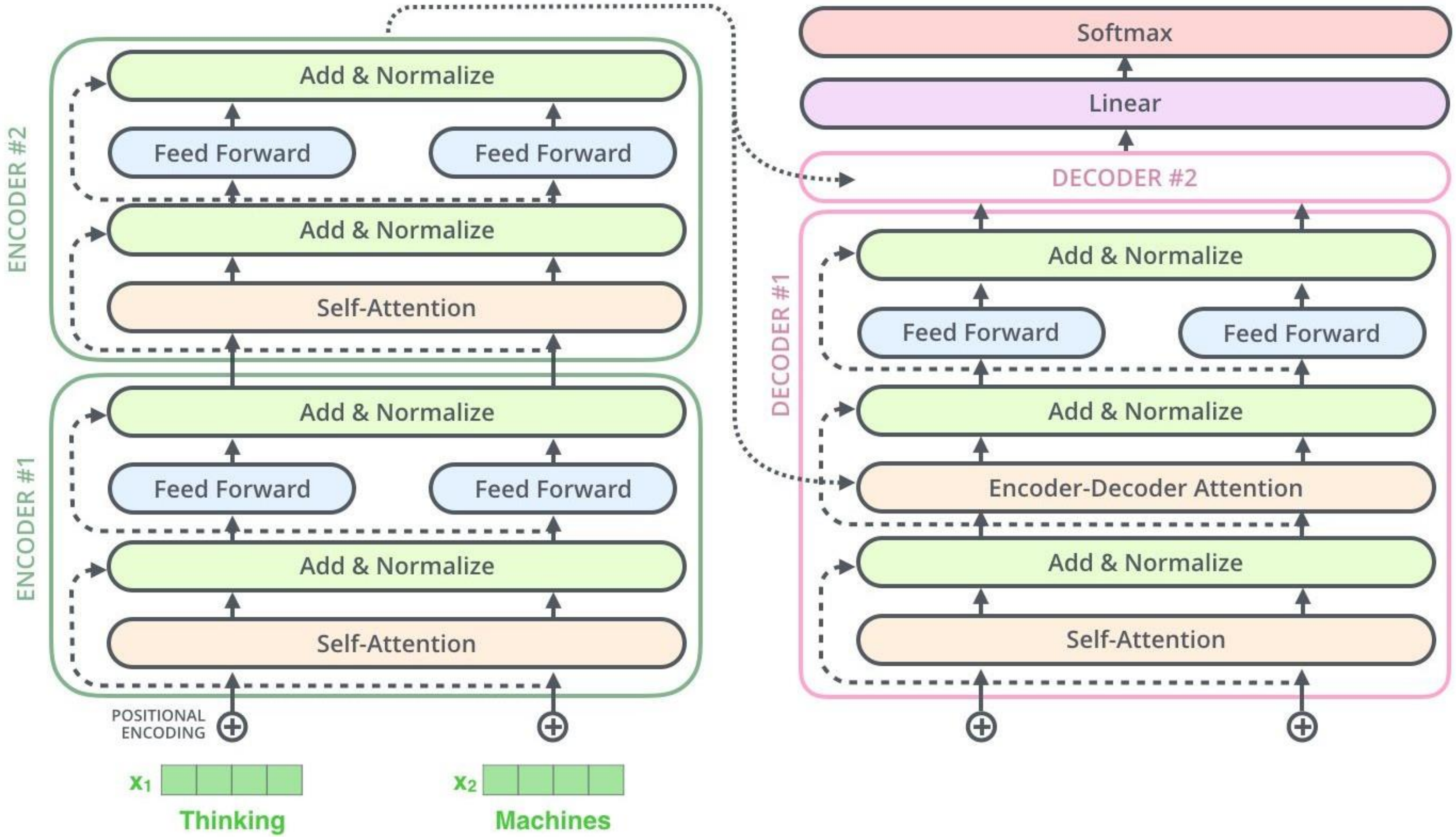


STS: Semantic Textual Similarity

score	2 example sentences	explanation
5	*The bird is bathing in the sink.* *Birdie is washing itself in the water basin.*	The two sentences are completely equivalent, as they mean the same thing.
4	*Two boys on a couch are playing video games.* *Two boys are playing a video game.*	The two sentences are mostly equivalent, but some unimportant details diffe
3	*John said he is considered a witness but not a suspect.* *“He is not a suspect anymore.” John said.*	The two sentences are roughly equivalent, but some important information differs/missing.
2	*They flew out of the nest in groups.* *They flew into the nest together.*	The two sentences are not equivalent, but share some details.
1	*The woman is playing the violin.* *The young lady enjoys listening to the guitar.*	The two sentences are not equivalent, but are on the same topic.
0	*The black dog is running through the snow.* *A race car driver is driving his car through the mud.*	The two sentences are completely dissimilar.


"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/

pital and largest city of the u
ma. the county seat of oklah
ty ranks 27th among united
tion. the population grew fol
s, with the population estim
ed to 643,648 as of july 201
oklahoma city metropolitan
of 1,358,452,[9] and the
hawnee combined statistic
of 1,459,758 residents,[9]
oma's largest metropolitan a

city limits extend into cana

running man was c
variety"; a genre of
environment.[1] the
complete missions
race.[2] the show h
familiar reality-vari
games. it has garn
comeback program
of the program, aft
family outing in feb

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

"whe
"barro
englis
carryi

the wh
weigh
opera
heavi
were t
as su

county,[8] the c
cities in popula
the 2010 census
to have increas
as of 2015, the
had a populat
oklahoma city's
had a populati
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...

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
d 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 c
as such it is a second-class lever...

eaty of p

reaty of paris, also k
, was signed on 10 fe
doms of great britain,
ugal in a
france a

igning o
s' war, k
orth a
nning of
pe.[2] gr
h of the
g the w
ce's po
tionally,

== lemon

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

d for
roughout
has both
and rind
king. the
itric acid,
taste. the

== 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

greed to protect roman
rld...

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
erican theatre,[1] and marked
an era of british dominance ou
eat britain and france each ret
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.) Osb
species of small evergreen tree
flowering plant family rutaceae
The tree's ellipsoidal yellow frui
culinary and non-culinary purpos
throughout the world, primarily
which has both culinary and cle
The juice of the lemon is about
citric acid, with a ph of around
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

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.

erdum
eget

e lacinia at

fficitur

em non,

a sodales in

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

- Removed lines that didn't end in a terminal punctuation mark.
- Language classifier to retain only English text
- Removed texts which look like placeholder texts
- Removed anything which look like code
- Removed duplicated texts

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);
   }
}
```

c4 (Manual download)

Contents

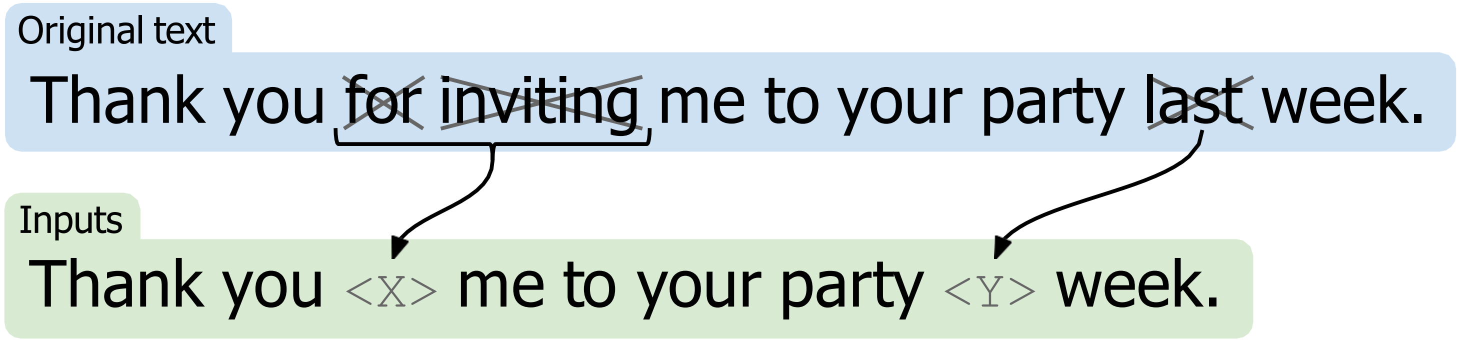
A colossal, cleaned version of Common Crawl's web crawl corpus.

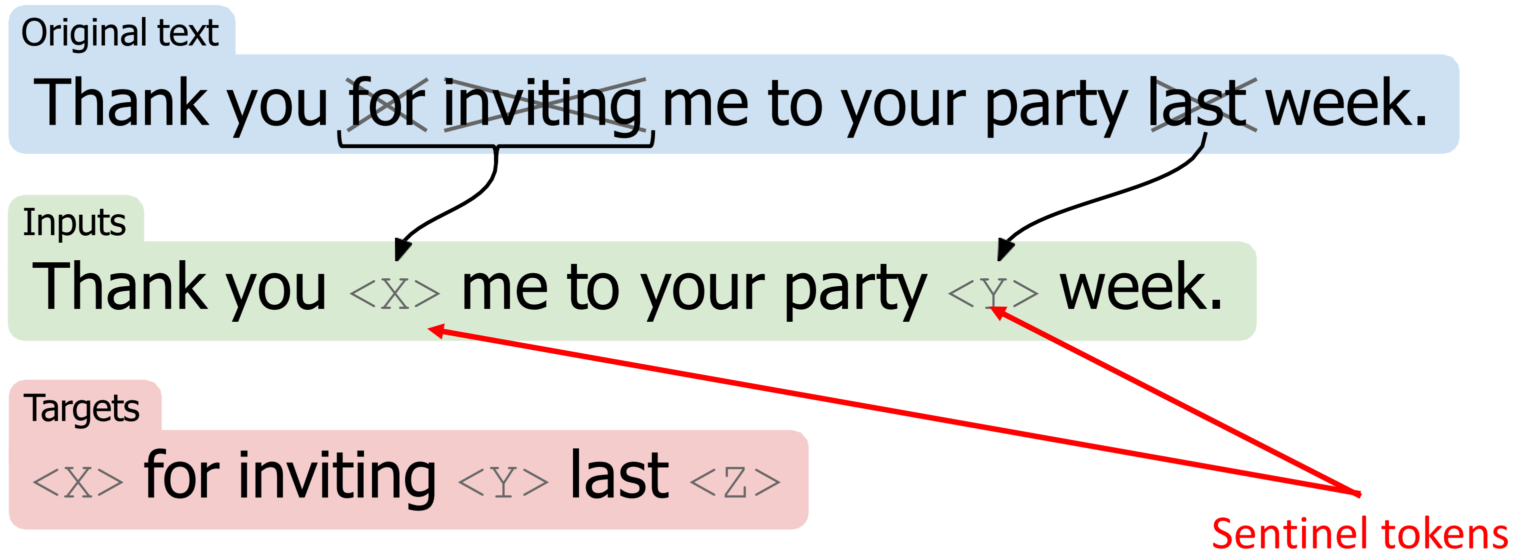
Original text

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>
Sentinel tokens

Pretrain

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

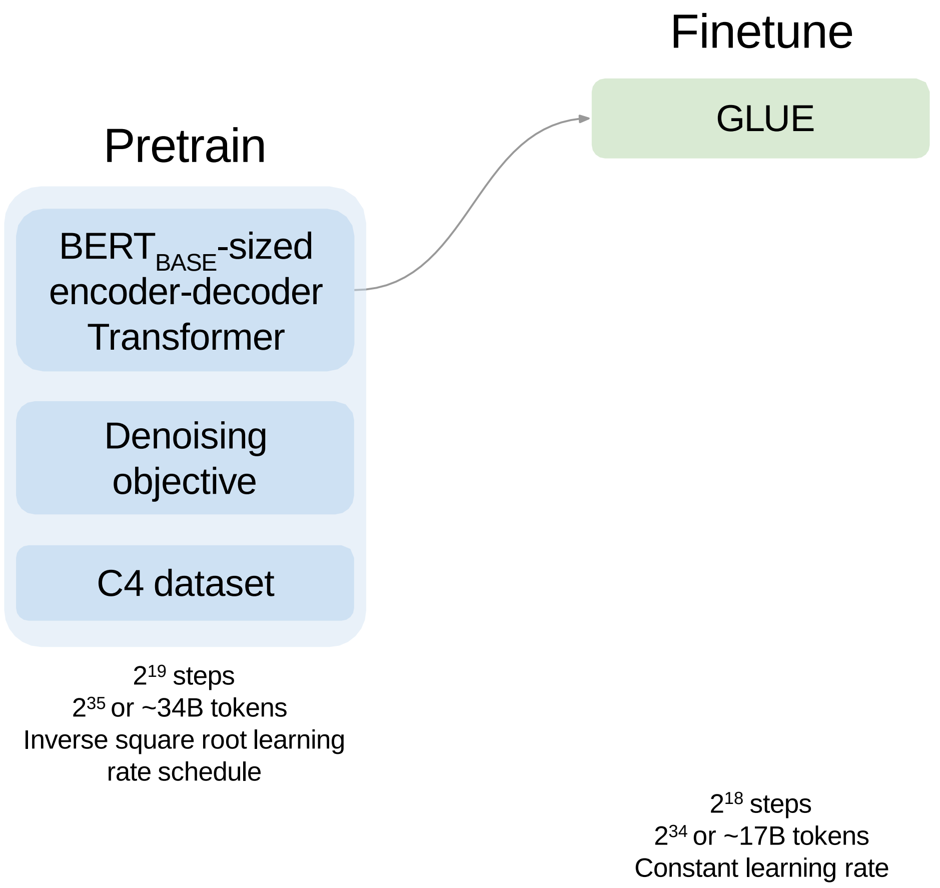
Denoising objective

C4 dataset

2^{19} steps

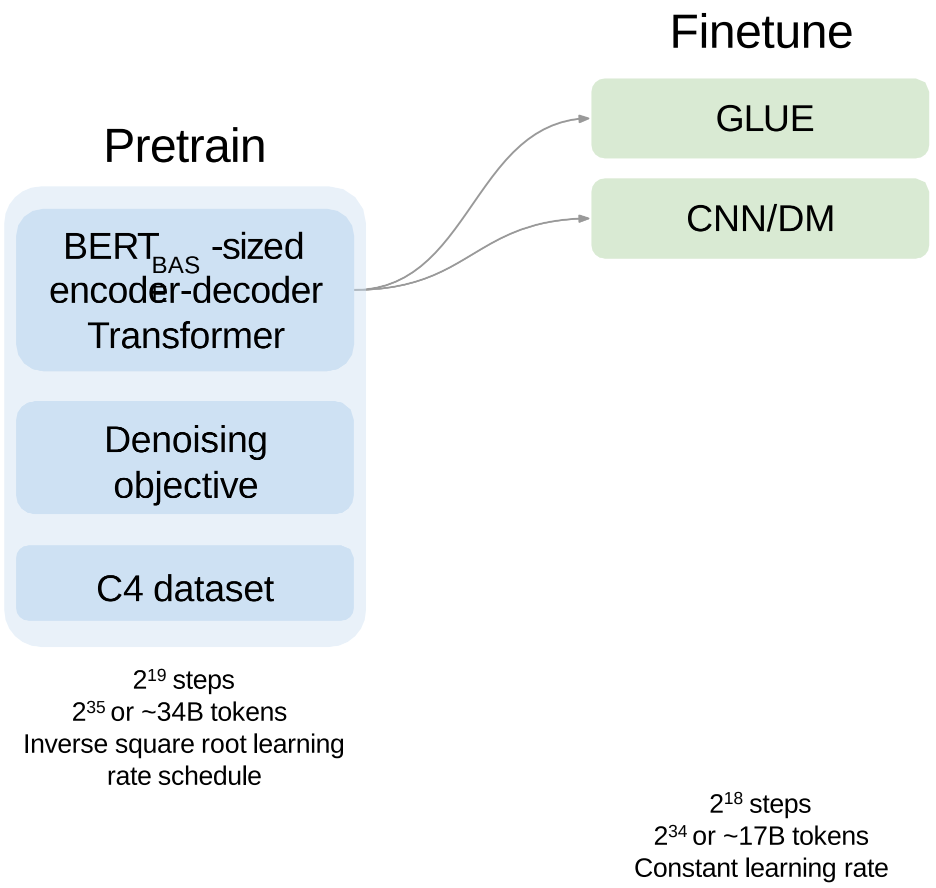
2^{35} or ~34B tokens

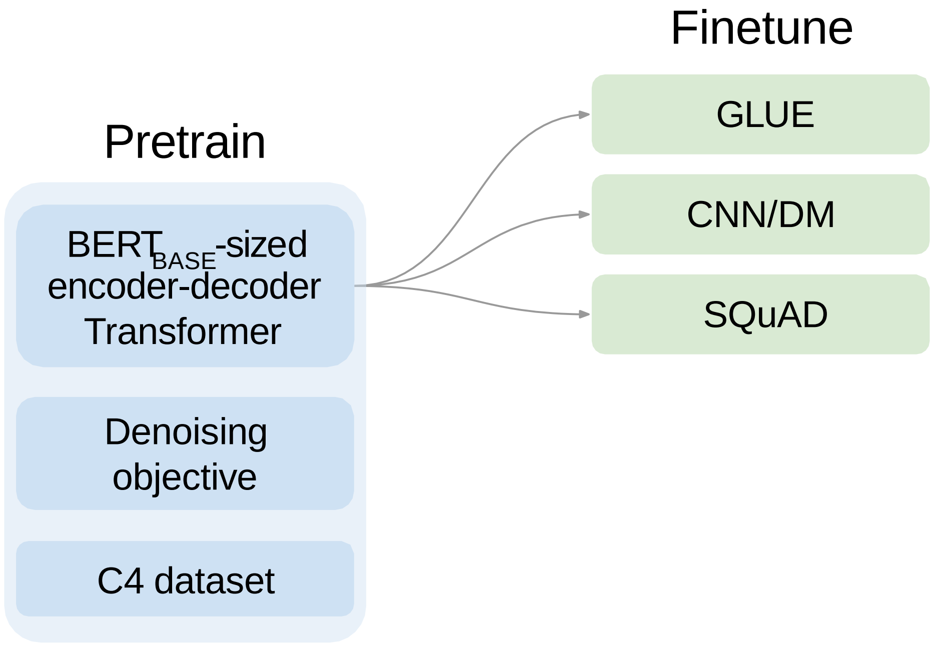
Inverse square root learning rate schedule


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

GLUE Benchmark

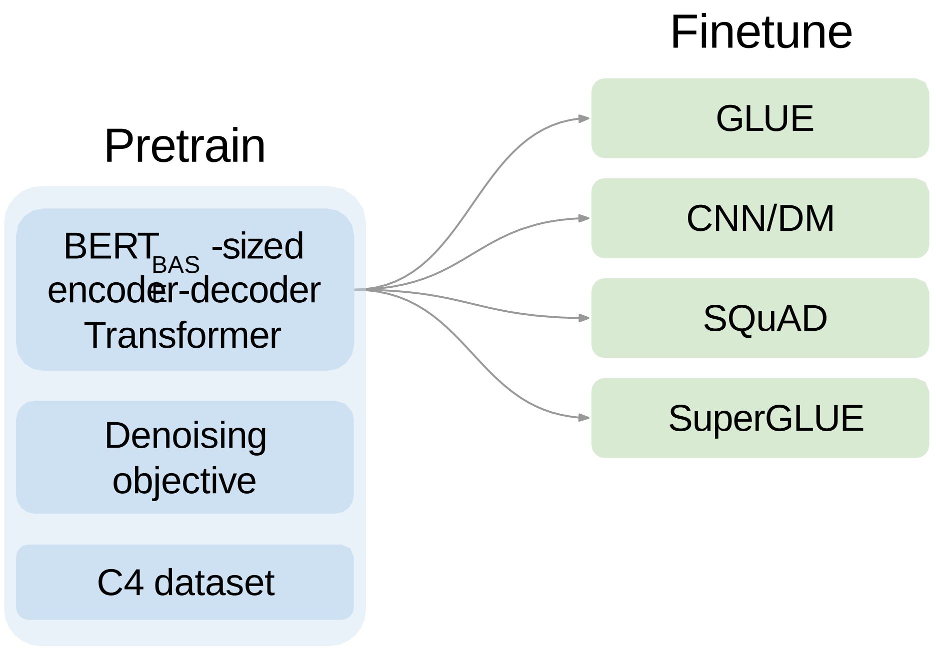
Dataset	Description	Data example	Metric
CoLA	Is the sentence grammatical or ungrammatical?	"This building is than that one." = **Ungrammatical**	Matthews
SST-2	Is the movie review positive, negative, or neutral?	"The movie is funny , smart , visually inventive , and most of all , alive ." = **.93056 (Very Positive)**	Accuracy
MRPC	Is the sentence B a paraphrase of sentence A?	A) "Yesterday , Taiwan reported 35 new infections , bringing the total number of cases to 418 ." B) "The island reported another 35 probable cases yesterday , taking its total to 418 ." = **A Paraphrase**	Accuracy / F1
STS-B	How similar are sentences A and B?	A) "Elephants are walking down a trail." B) "A herd of elephants are walking along a trail." = **4.6 (Very Similar)**	Pearson / Spearman
QQP	Are the two questions similar?	A) "How can I increase the speed of my internet connection while using a VPN?" B) "How can Internet speed be increased by hacking through DNS?" = **Not Similar**	Accuracy / F1
MNLI-mm	Does sentence A entail or contradict sentence B?	A) "Tourist Information offices can be very helpful." B) "Tourist Information offices are never of any help." = **Contradiction**	Accuracy
QNLI	Does sentence B contain the answer to the question in sentence A?	A) "What is essential for the mating of the elements that create radio waves?" B) "Antennas are required by any radio receiver or transmitter to couple its electrical connection to the electromagnetic field." = **Answerable**	Accuracy
RTE	Does sentence A entail sentence B?	A) "In 2003, Yunus brought the microcredit revolution to the streets of Bangladesh to support more than 50,000 beggars, whom the Grameen Bank respectfully calls Struggling Members." B) "Yunus supported more than 50,000 Struggling Members." = **Entailed**	Accuracy
WNLI	Sentence B replaces sentence A's ambiguous pronoun with one of the nouns - is this the correct noun?	A) "Lily spoke to Donna, breaking her concentration." B) "Lily spoke to Donna, breaking Lily's concentration." = **Incorrect Referent**	Accuracy


Finetune
Pretrain
GLUE
CNN/DM
BERT BAS -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



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

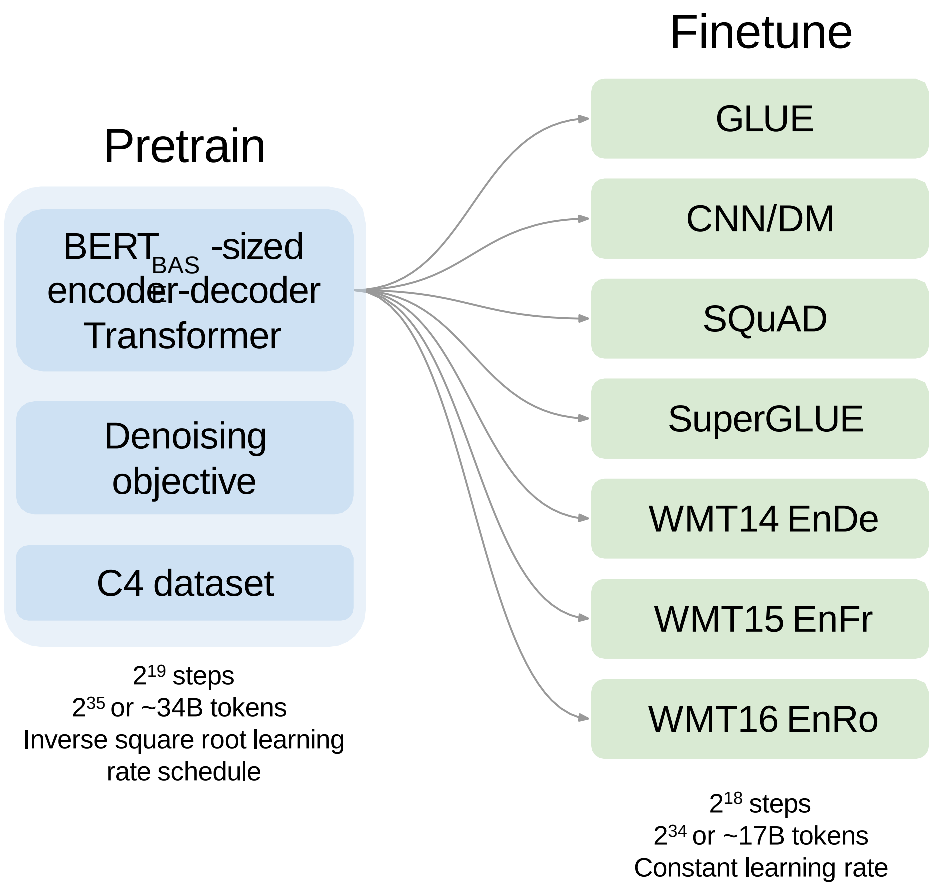


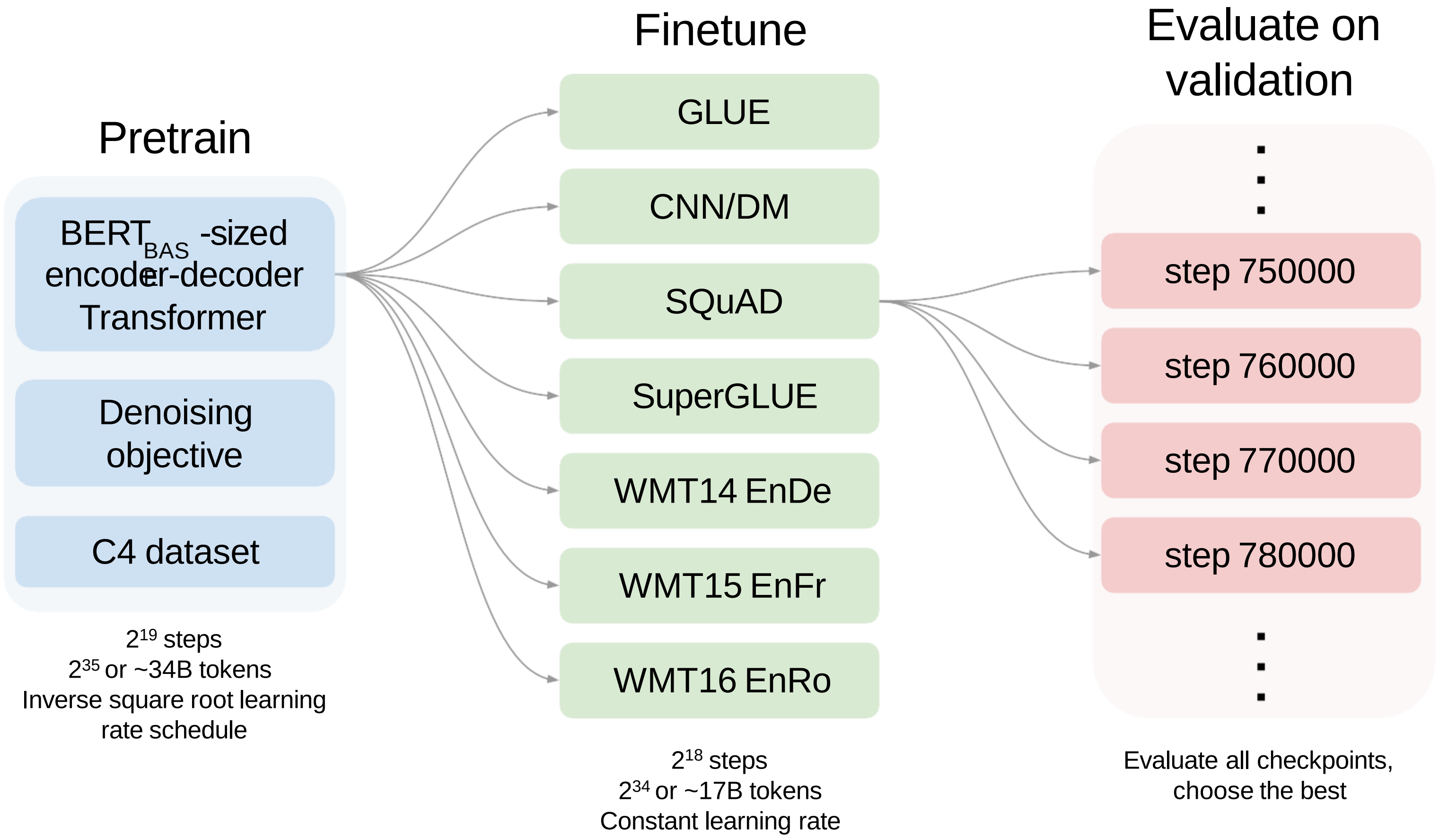
2^{19} steps
2^{35} or ~34B tokens
Inverse square root learning rate schedule

2^{18} steps
2^{34} or ~17B tokens
Constant learning rate

SuperGLUE Tasks

Name	Identifier	Download	More Info	Metric
Broadcoverage Diagnostics	AX-b			Matthew's Corr
CommitmentBank	CB			Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA			Accuracy
Multi-Sentence Reading Comprehension	MultiRC			F1a / EM
Recognizing Textual Entailment	RTE			Accuracy
Words in Context	WiC			Accuracy
The Winograd Schema Challenge	WSC			Accuracy
BoolQ	BoolQ			Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD			F1 / Accuracy
Winogender Schema Diagnostics	AX-g			Gender Parity / Accuracy


Pretrain
BERT$_{\text{BAS}}$ -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


Pretrain
BERT BAS -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

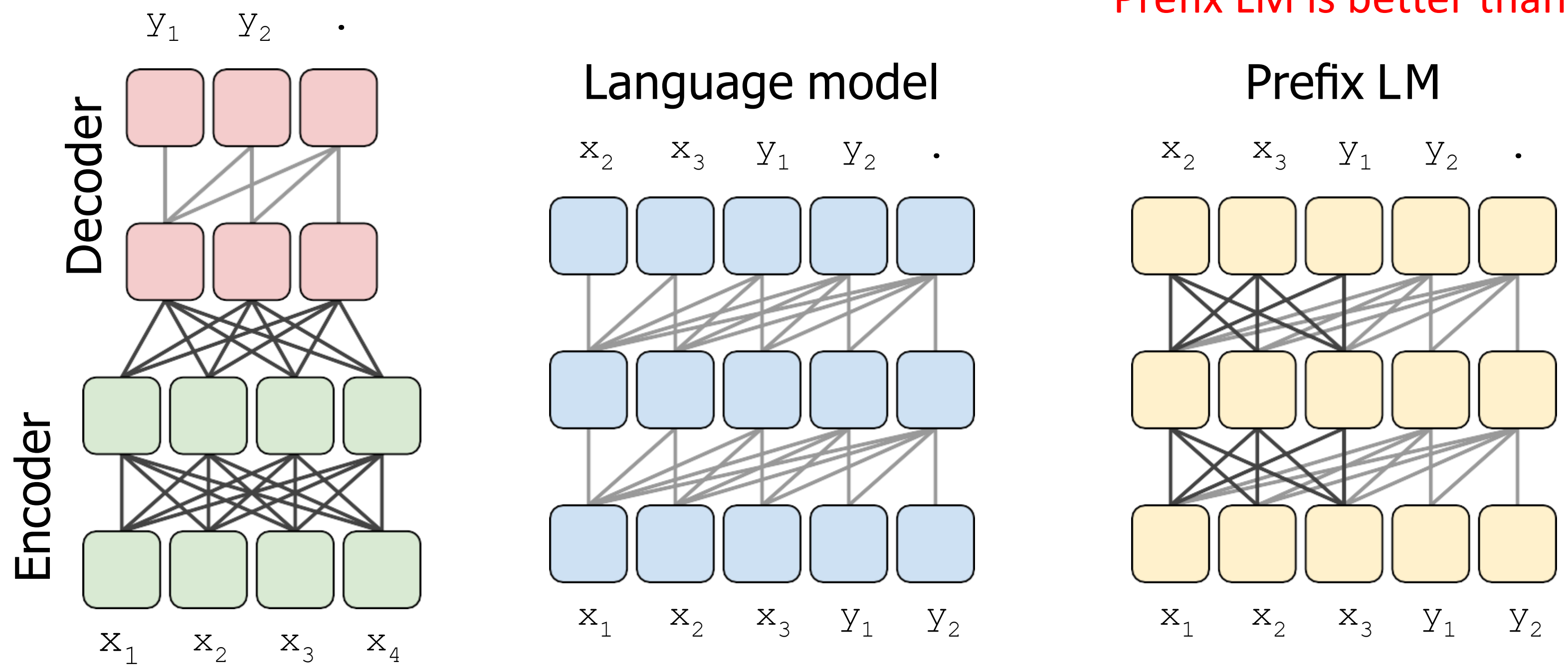
No pre-training is dramatically worse, except EnFr!

Disclaimer

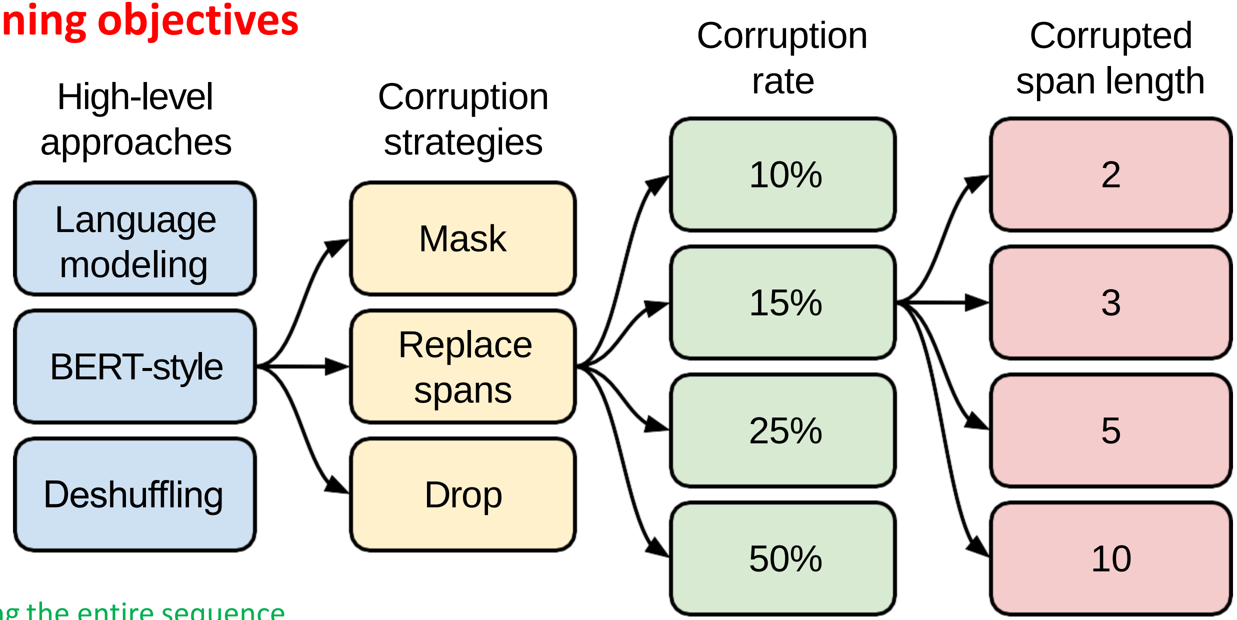
We will not tweak any hyper-parameter in the rest of the slides

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

Prefix LM is better than LM



Different pre-training objectives



First two involve predicting the entire sequence
Last two predict only the masked/dropped tokens => lower petraining cost

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**

Different pre-training datasets

Smashwords

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

Different pre-training datasets

Please enable JavaScript to use our site.

Please enable JavaScript to use our site.

Wiki+TBC for SGLUE, better performance => SGLUE has a reading comprehension task, MultiRC

MultiRC (Multi-Sentence Reading Comprehension) is a dataset of short paragraphs and multi-sentence questions, i.e., questions that can be answered by combining information from multiple sentences of the paragraph

Sent 1: The hijackers attacked at 9:28.
Sent 2: While traveling 35,000 feet above eastern Ohio, United 93 suddenly dropped 700 feet.
Sent 3: Eleven seconds into the descent, the FAA's air traffic control center in Cleveland received the first of two radio transmissions from the aircraft.
Sent 4: During the first broadcast, the captain or first officer could be heard declaring "Mayday" amid the sounds of a physical struggle in the cockpit.
Sent 5: The second radio transmission, 35 seconds later, indicated that the fight was continuing.
Sent 6: The captain or first officer could be heard shouting: "Hey get out of here-get out of here-get out of here."
Sent 7: On the morning of 9/11, there were only 37 passengers on United 93-33 in addition to the 4 hijackers.
Sent 8: This was below the norm for Tuesday mornings during the summer of 2001.
Sent 9: But there is no evidence that the hijackers manipulated passenger levels or purchased additional seats to facilitate their operation.
Sent 10: The terrorists who hijacked three other commercial flights on 9/11 operated in five-man teams.
Sent 11: They initiated their cockpit takeover within 30 minutes of takeoff.
Sent 12: On Flight 93, however, the takeover took place 46 minutes after takeoff and there were only four hijackers.

Question: Which two factors were different between the three other hijacked planes and United 93?
the day of the takeover
A) The amount of time that passed before the takeover started
B)* United 93 took longer and had less hijackers
C) The airline operating the planes
D) The weather and fuel used by the airplane
E) The navigation system used by the planes

Reasoning needed: Discourse relation (contrast)
One needs to identify that the discourse marker *however* in Sent 12 indicates a contrast relation between Flight 93 and the flights mentioned in Sent 10. Also, *only* in Sent 12 indicates that the number of hijackers were fewer than in the contrasted other flights.

Dataset	Size	GLUE	CNNDM	S				
★ C4	745GB	83.28	**19.24**					
C4, unfiltered	6.1TB	81.46	19.14					
RealNews-like	35GB	**83.83**	**19.23**					
WebText-like	17GB	**84.03**	**19.31**					
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**

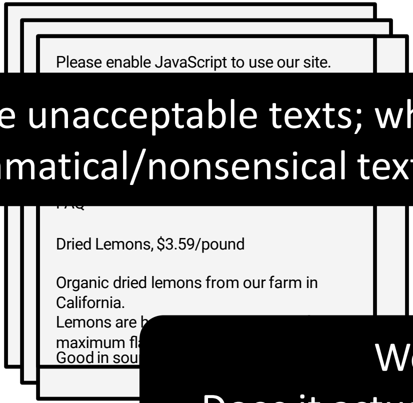
Much worse on CoLA

Much better on ReCoRD

Order of magnitude smaller

Much better on MultiRC

Different pre-training datasets



Wikipedia has very little unacceptable texts; where C4 has many ungrammatical/nonsensical texts

words

We also see gains on smaller datasets
Does it actually hurt you to pretrain on a smaller 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**

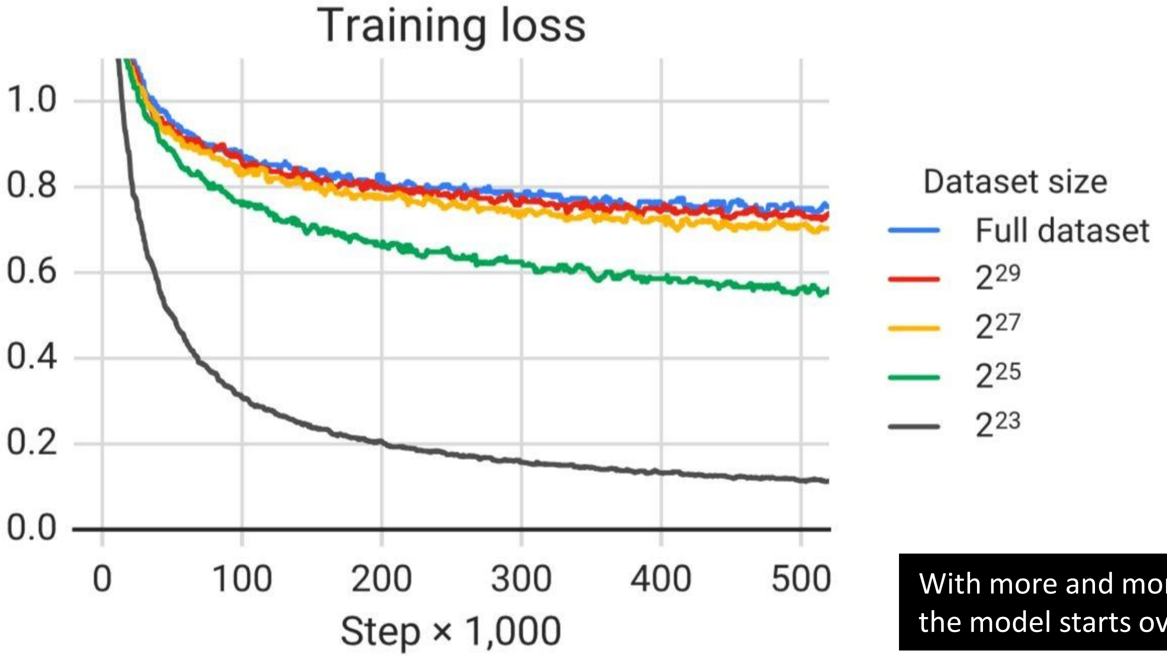
Much worse on CoLA

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



With more and more repetition, the model starts overfitting

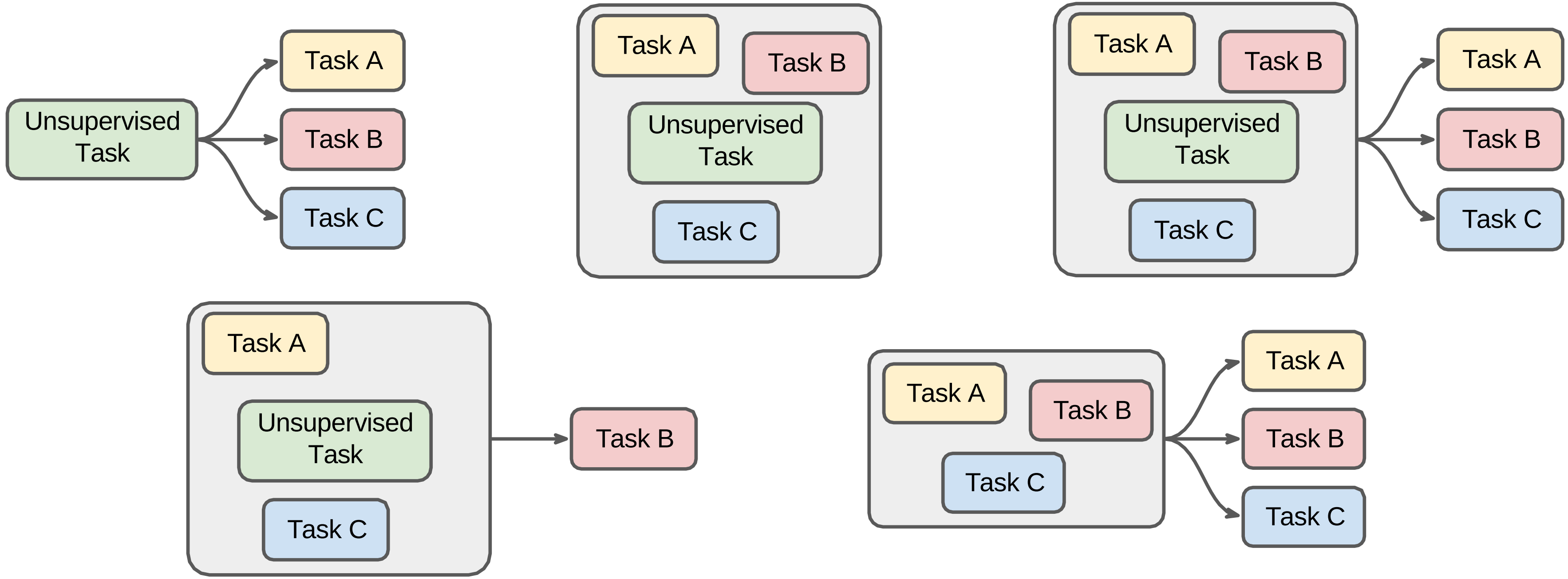
Mixing strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline (pre-train/fine-tine)	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Equal	76.13	19.02	76.51	63.37	23.89	34.31	26.78
Examples-proportional, $K = 2^{16}$	80.45	19.04	77.25	69.95	24.35	34.99	27.10
Examples-proportional, $K = 2^{17}$	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

Examples-proportional Specifically, if the number of examples in each of our N task's data sets is $e_n, n \in \{1, \ldots, N\}$ then we set probability of sampling an example from the mth task during training to $r_m = \min(e_m, K)/\sum \min(e_n, K)$ where K is the artificial data set size limit.

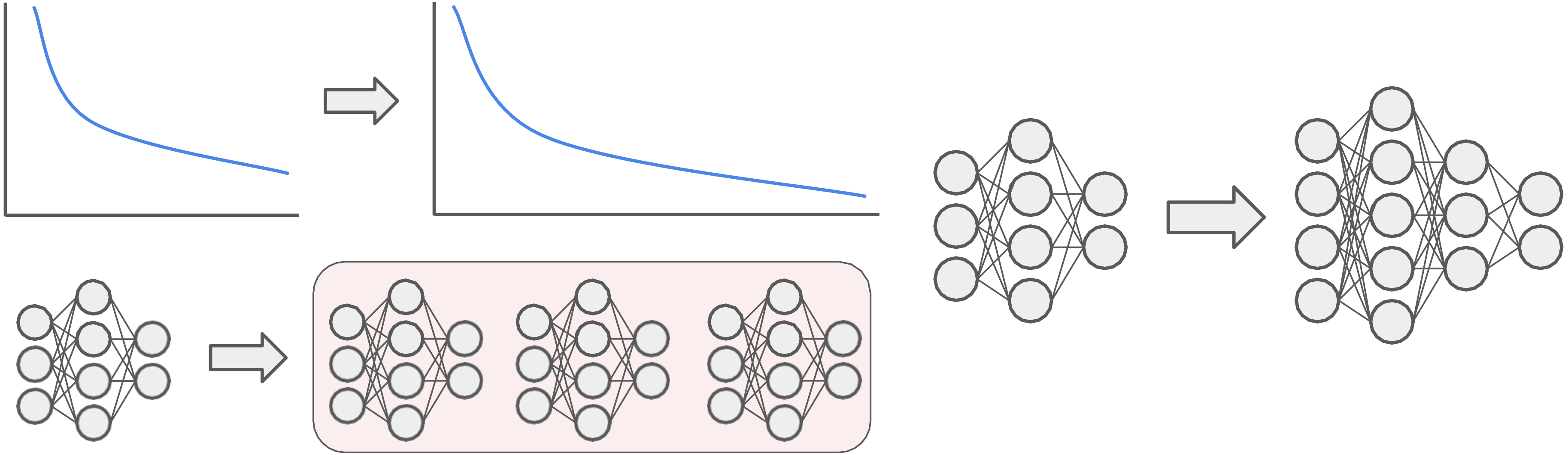
- You can get close to the performance of the baseline if we do the mixing strategy right.
- You do tend to sacrifice some performance when doing multitask on at least some of the tasks

Comparing Multitask Learning with Fine-tuning

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 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



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

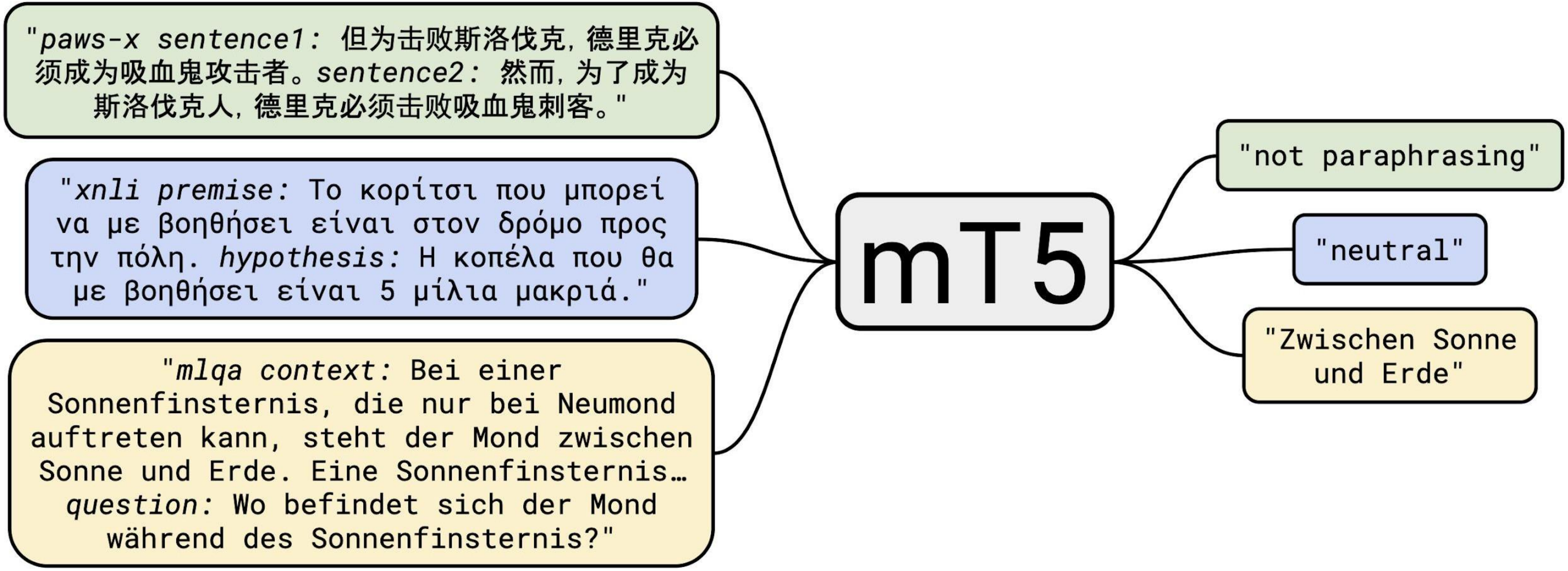
Why it does not perform well for MT?

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

What about all of the other languages?


"paws-x sentence1: 但为击败斯洛伐克，德里克必须成为吸血鬼攻击者。sentence2: 然而，为了成为斯洛伐克人，德里克必须击败吸血鬼刺客。"
"xnli premise: Το κορίτσι που μπορεί να με βοηθήσει είναι στον δρόμο προς την πόλη. hypothesis: Η κοπέλα που θα με βοηθήσει είναι 5 μίλια μακριά."
"mlqa context: Bei einer Sonnenfinsternis, die nur bei Neumond auftreten kann, steht der Mond zwischen Sonne und Erde. Eine Sonnenfinsternis… question: Wo befindet sich der Mond während des Sonnenfinsternis?"
mT5
"not paraphrasing"
"neutral"
"Zwischen Sonne und Erde"

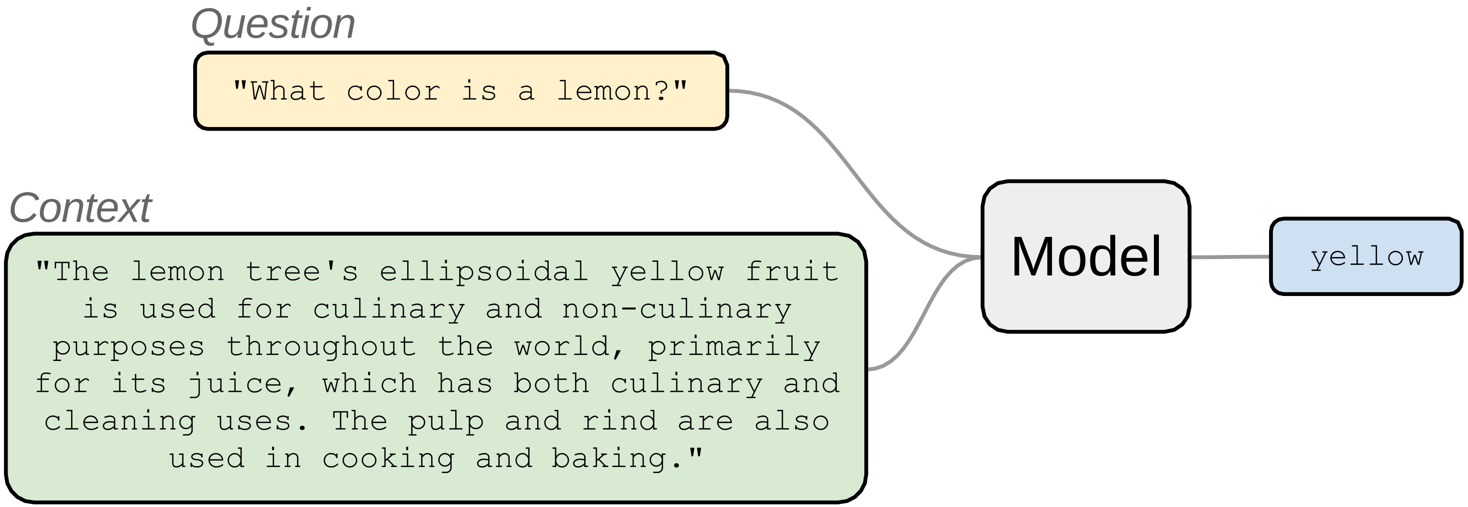
c4/multilingual

- **Config description**: Multilingual C4 (mC4) has 101 languages and is generated from 71 Common Crawl dumps.
- **Download size**: `22.74 MiB`
- **Dataset size**: `26.76 TiB`

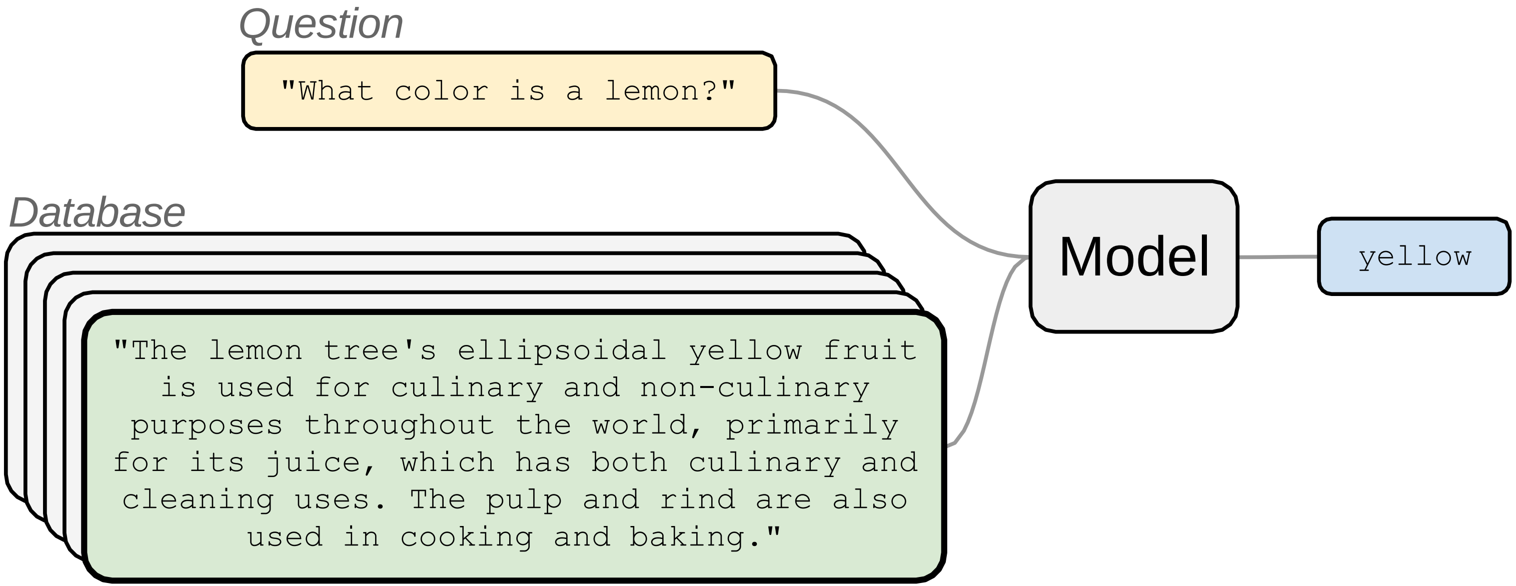
Afrikaans, Albanian, Amharic, Arabic, Armenian, Azerbaijani, Basque, Belarusian, Bengali, Bulgarian, Burmese, Catalan, Cebuano, Chichewa, Chinese, Corsican, Czech, Danish, Dutch, English, Esperanto, Estonian, Filipino, Finnish, French, Galician, Georgian, German, Greek, Gujarati, Haitian Creole, Hausa, Hawaiian, Hebrew, Hindi, Hmong, Hungarian, Icelandic, Igbo, Indonesian, Irish, Italian, Japanese, Javanese, Kannada, Kazakh, Khmer, Korean, Kurdish, Kyrgyz, Lao, Latin, Latvian, Lithuanian, Luxembourgish, Macedonian, Malagasy, Malay, Malayalam, Maltese, Maori, Marathi, Mongolian, Nepali, Norwegian, Pashto, Persian, Polish, Portuguese, Punjabi, Romanian, Russian, Samoan, Scottish Gaelic, Serbian, Shona, Sindhi, Sinhala, Slovak, Slovenian, Somali, Sotho, Spanish, Sundanese, Swahili, Swedish, Tajik, Tamil, Telugu, Thai, Turkish, Ukrainian, Urdu, Uzbek, Vietnamese, Welsh, West Frisian, Xhosa, Yiddish, Yoruba, Zulu.

How much knowledge does a language model pick up during pre-training?

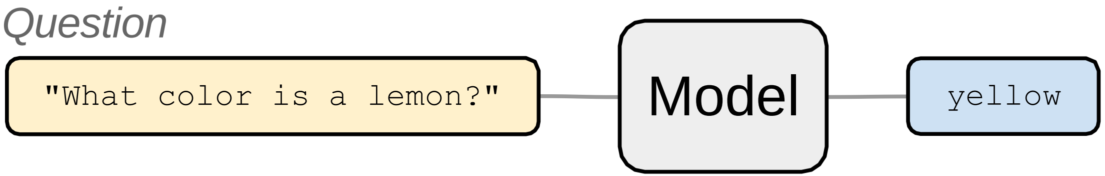
Reading Comprehension

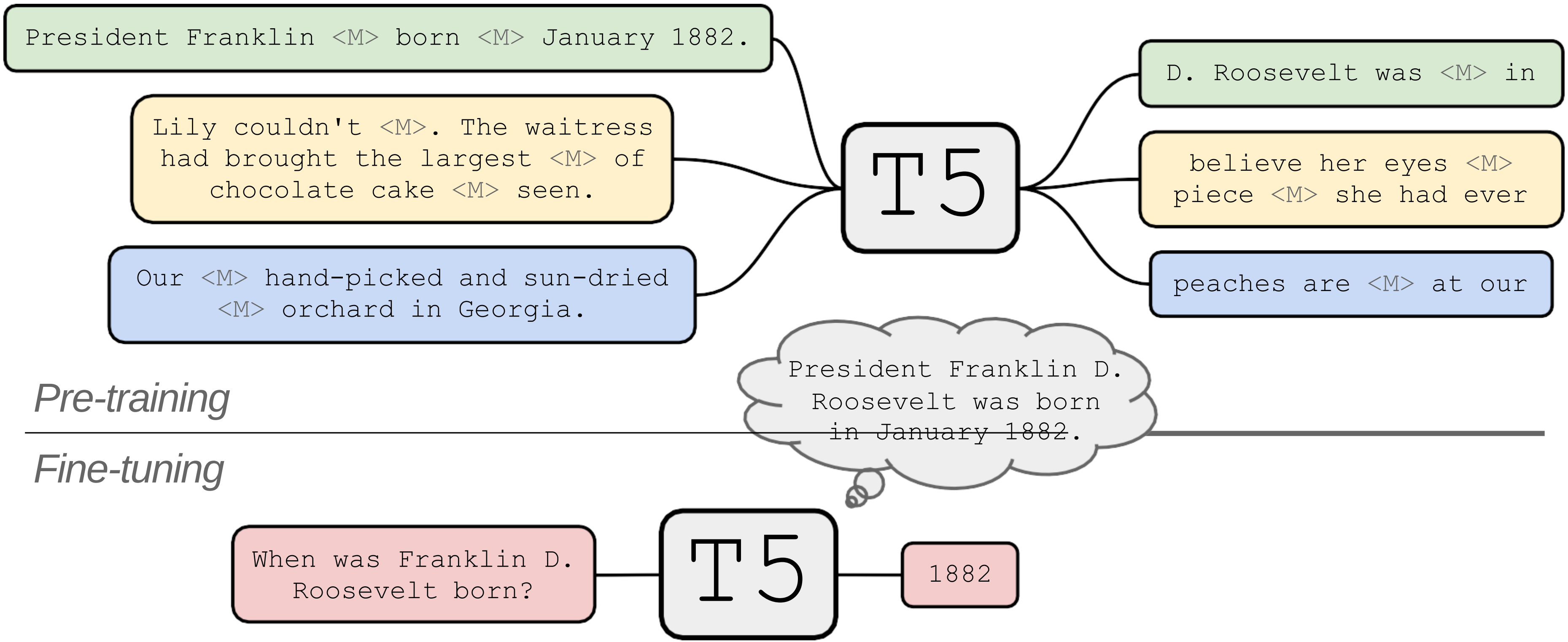


Open-Domain Question Answering



Closed-Book Question Answering




President Franklin <M> born <M> January 1882.
Lily couldn't <M>. The waitress had brought the largest <M> of chocolate cake <M> seen.
Our <M> hand-picked and sun-dried <M> orchard in Georgia.
T5
D. Roosevelt was <M> in
believe her eyes <M> piece <M> she had ever
peaches are <M> at our
Pre-training
President Franklin D. Roosevelt was born in January 1882.
Fine-tuning
When was Franklin D. Roosevelt born?
T5
1882

	NQ	WQ	TQA
Open-domain SoTA	41.5	42.4	57.9
T5.1.1-Base	25.7	28.2	24.2
T5.1.1-Large	27.3	29.5	28.5
T5.1.1-XL	29.5	32.4	36.0
T5.1.1-XXL	32.8	35.6	42.9