T5 and large language models: The good, the bad, and the ugly

Colin Raffel

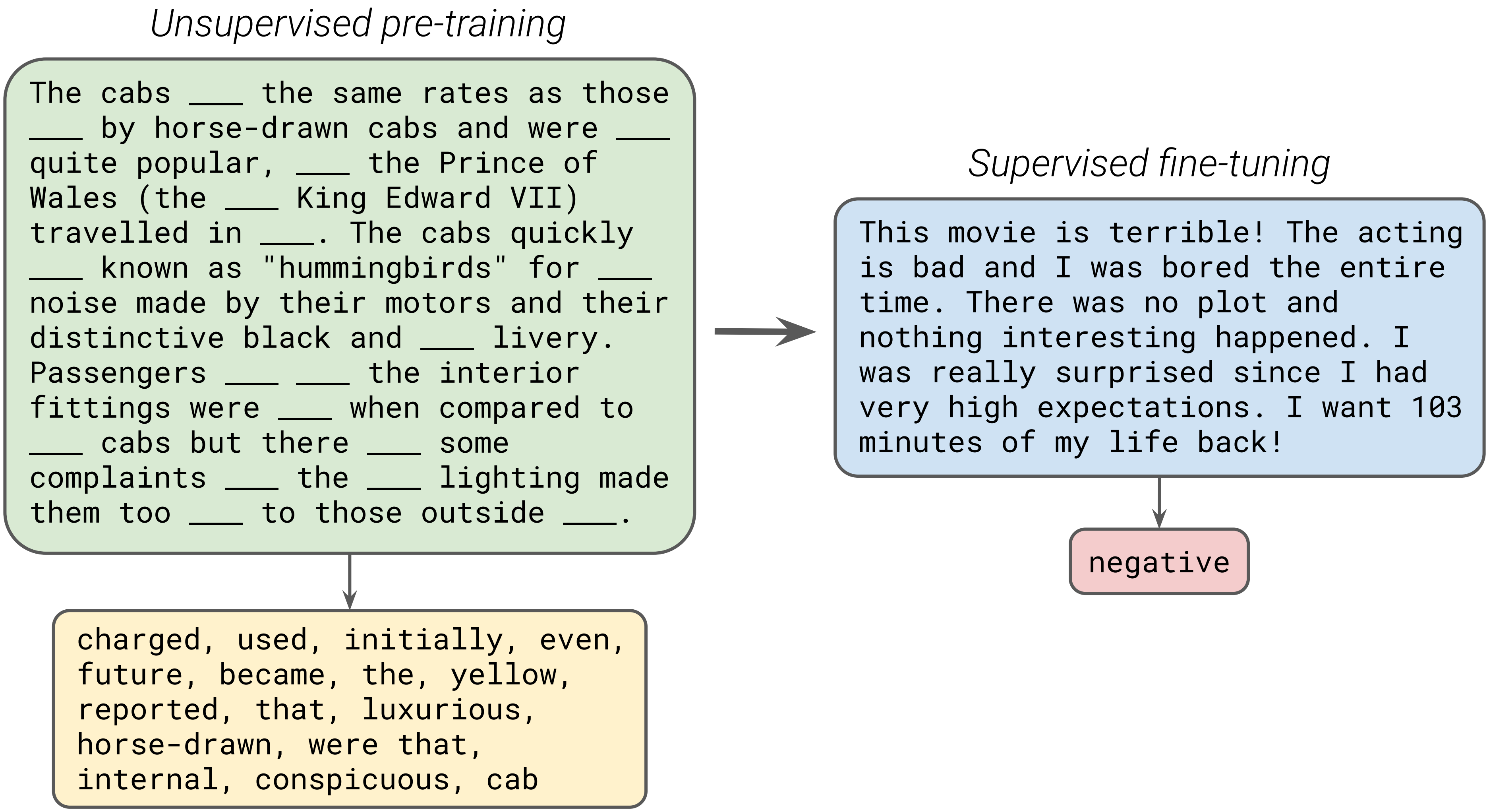
Which transfer learning methods work best, and what happens when we scale them up?

What about non-English pre-trained models?

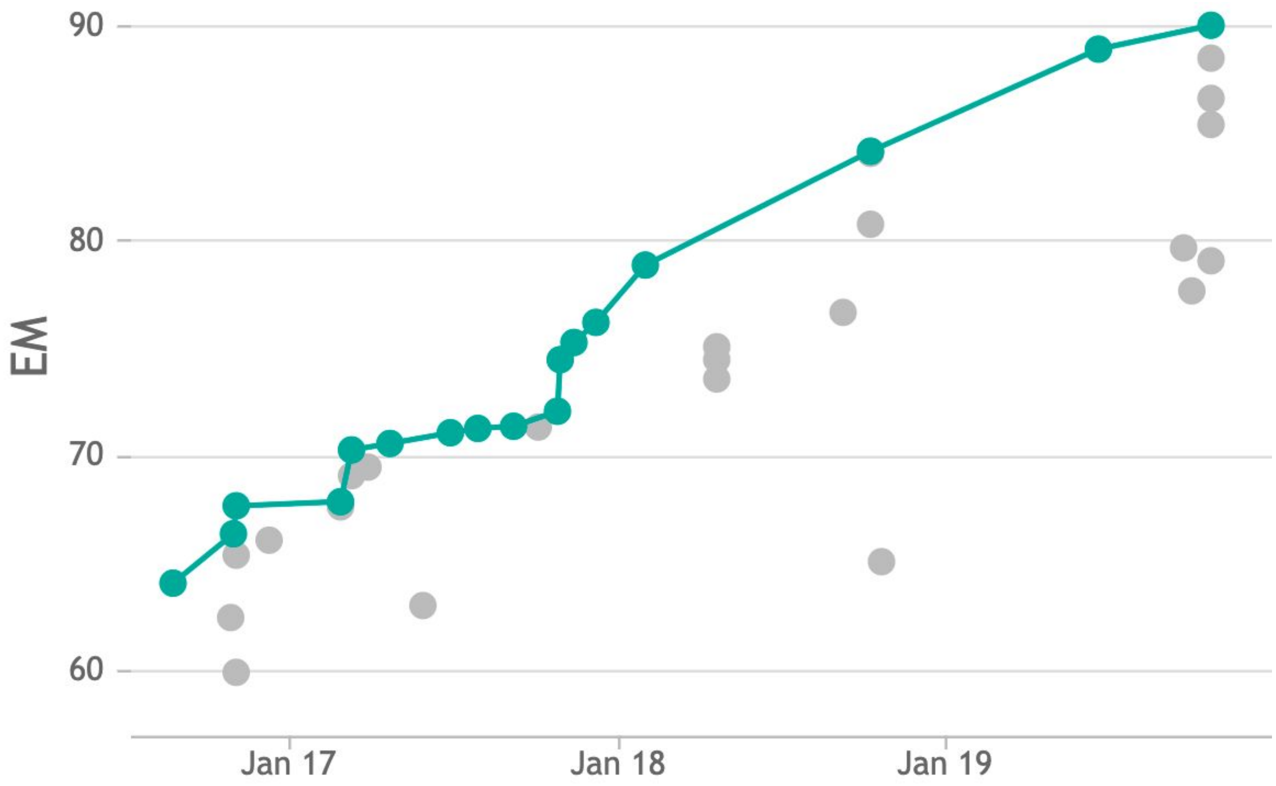
How much knowledge does the model learn during pre-training?

Does the model memorize data during pre-training?

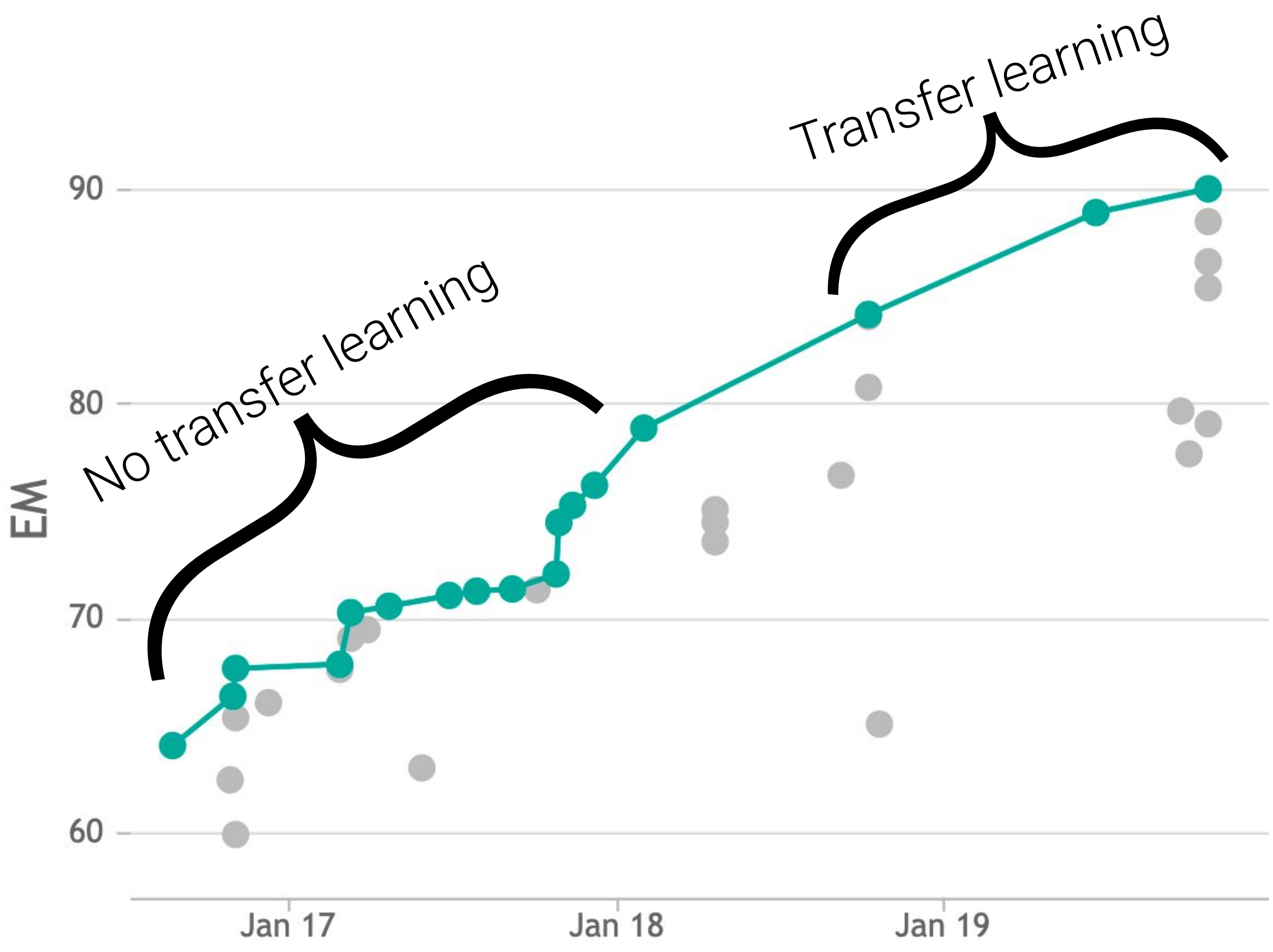
Which Transformer modifications work best?


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

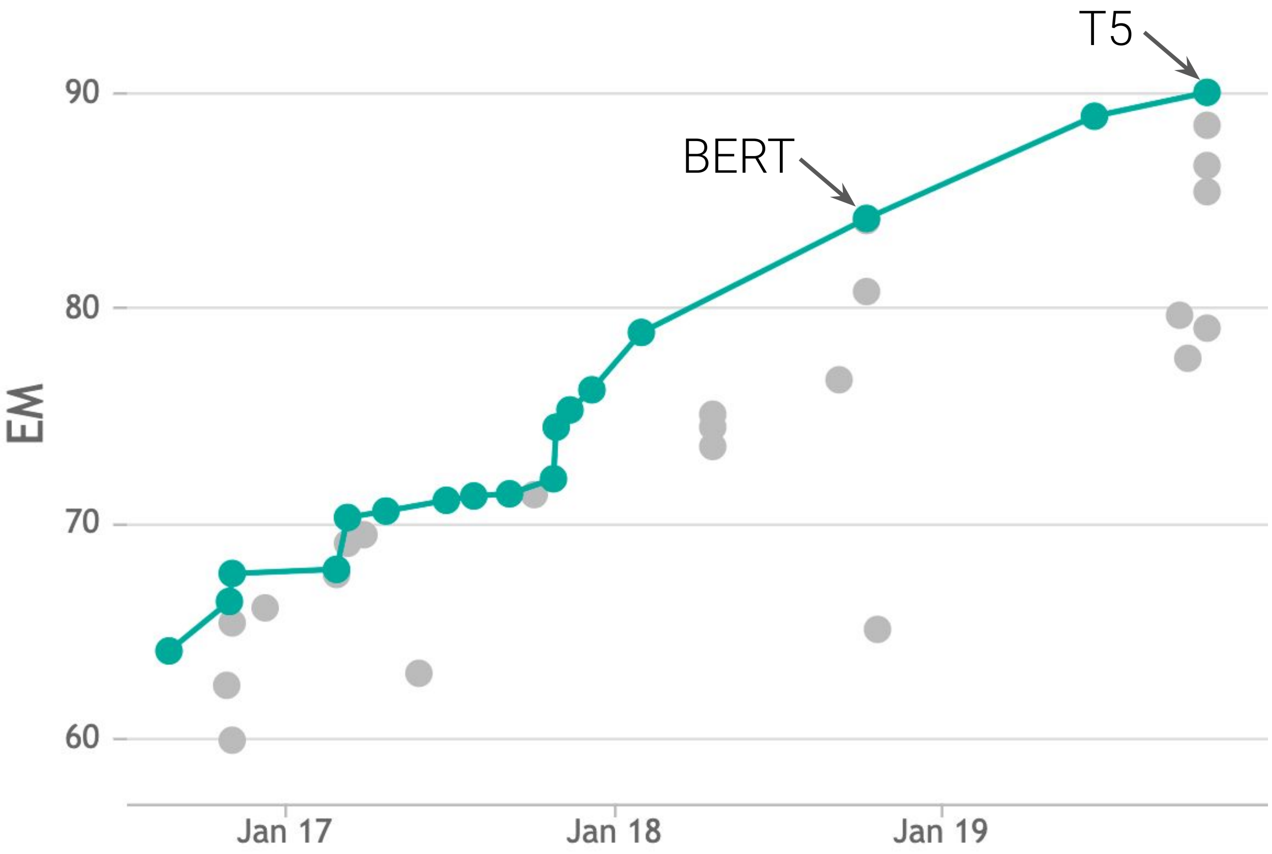
SQuAD Exact Match score (validation set)



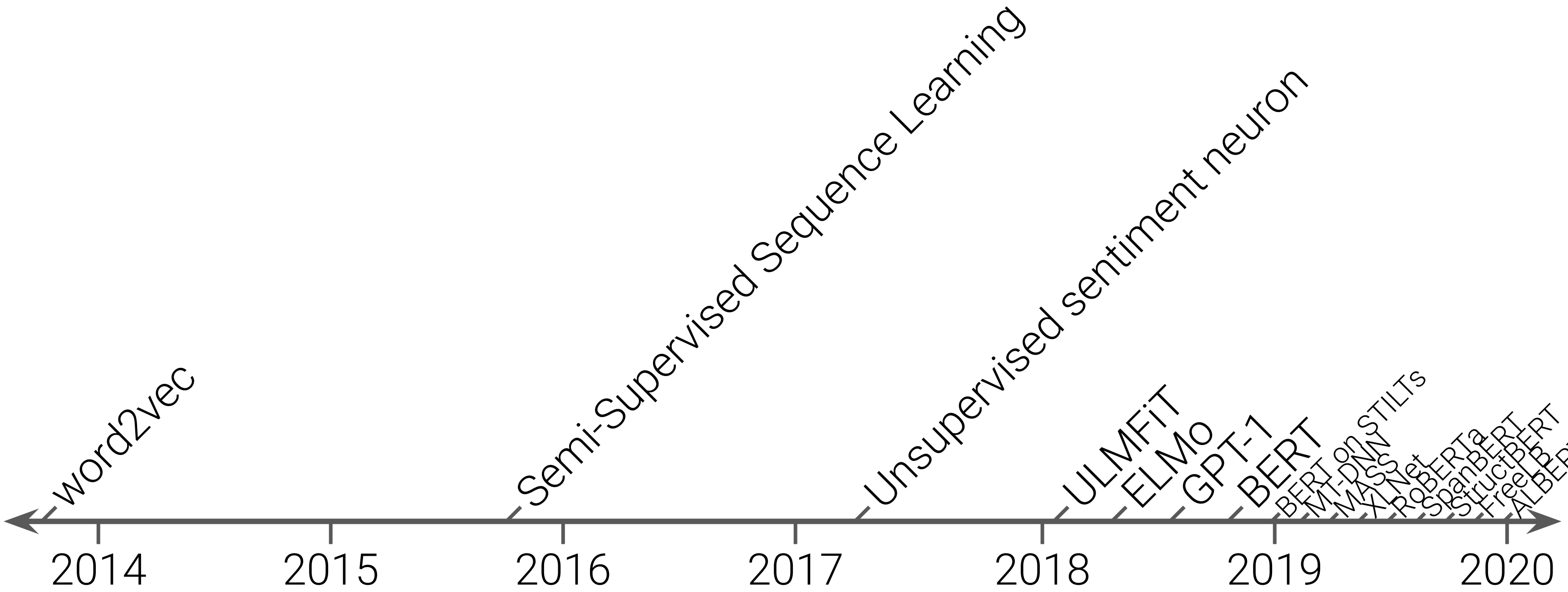
Source: https://paperswithcode.com/sota/question-answering-on-squad11-dev

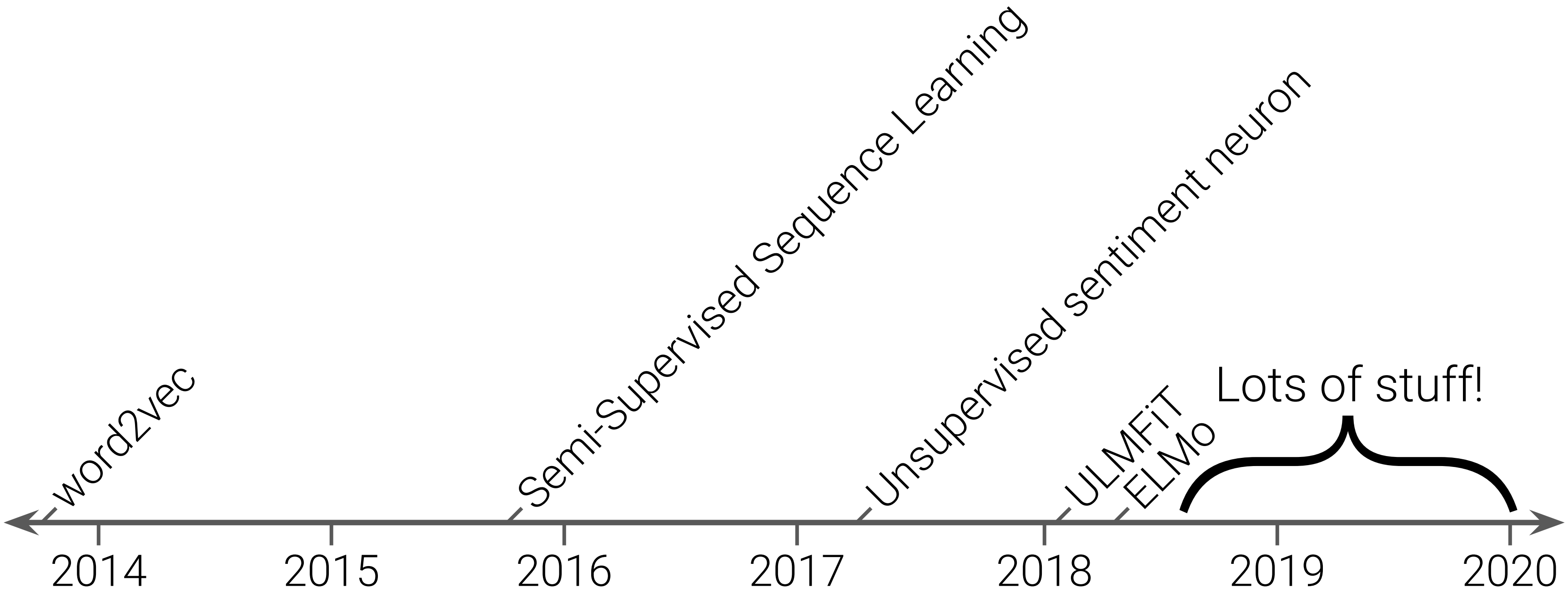


Source: https://paperswithcode.com/sota/question-answering-on-squad11-dev



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?*

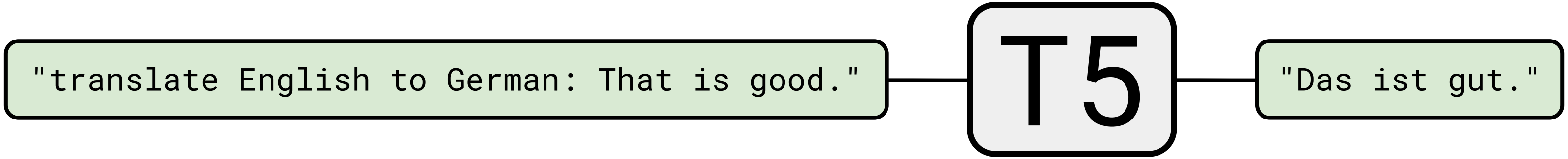
- Paper A proposes an unsupervised pre-training technique called "FancyLearn".
- Paper B proposes another pre-training technique called "FancierLearn" and achieves better results.
- Paper A pre-trains on **100 billion tokens** of unlabeled data.
- Paper B pre-trains on **200 billion tokens** of unlabeled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

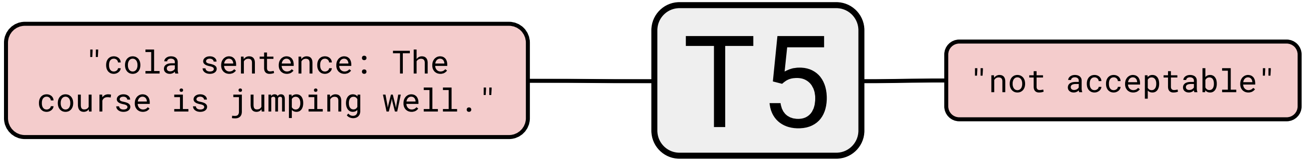
- Paper A proposes an unsupervised pre-training technique called "FancyLearn".
- Paper B proposes another pre-training technique called "FancierLearn" and achieves better results.
- Paper A uses the **Adam optimizer**.
- Paper B uses **SGD with momentum**.
- *Is FancierLearn better than FancyLearn?*

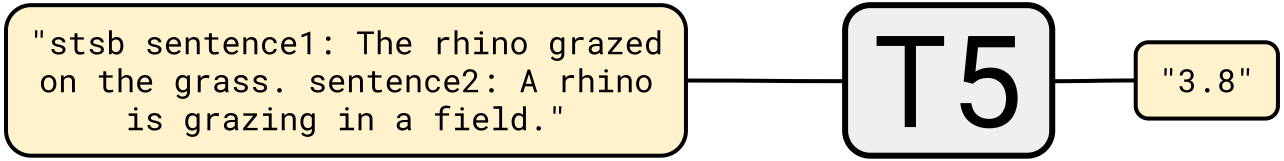
Given the current landscape of transfer learning for NLP, *what works best*? And how far can we push the tools we already have?

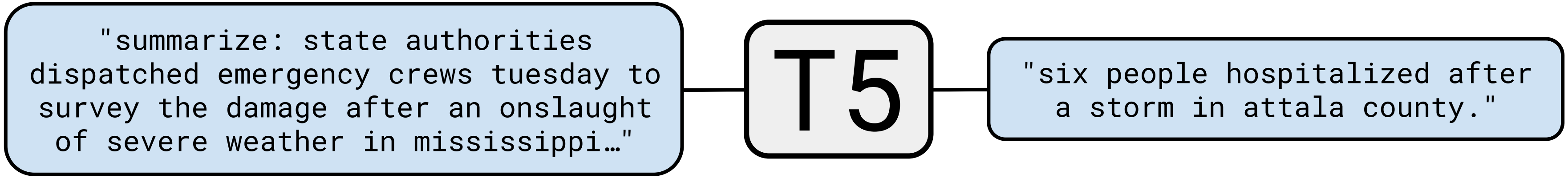
Text-to-Text
Transfer
Transformer

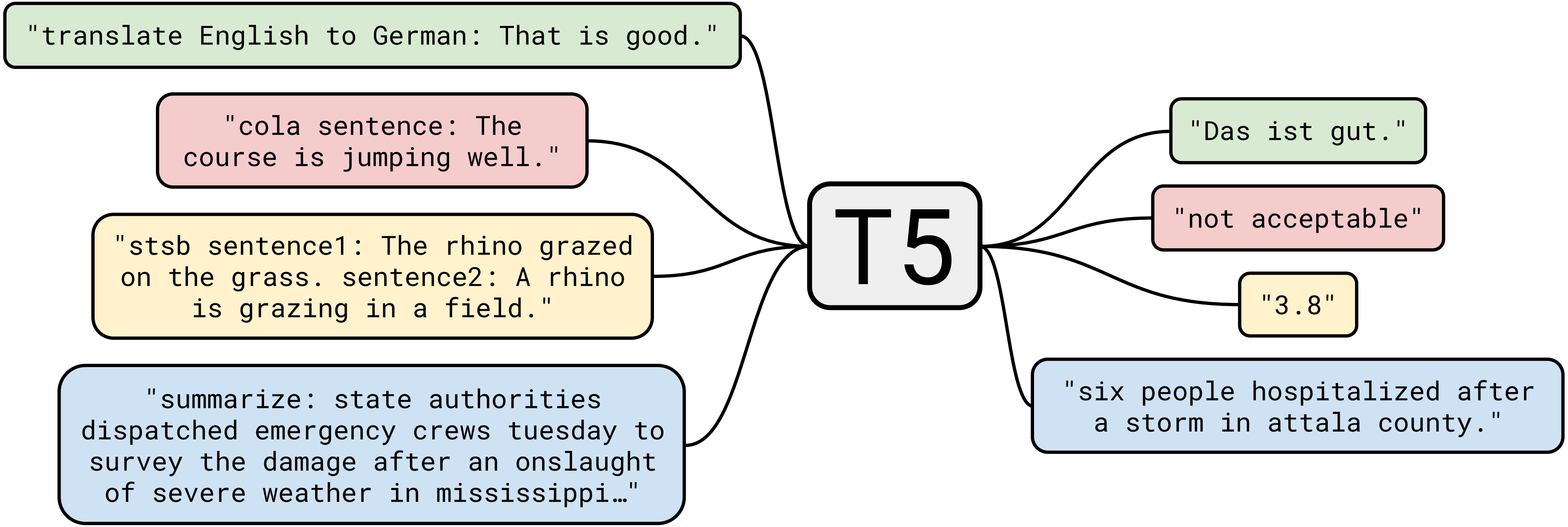
T5

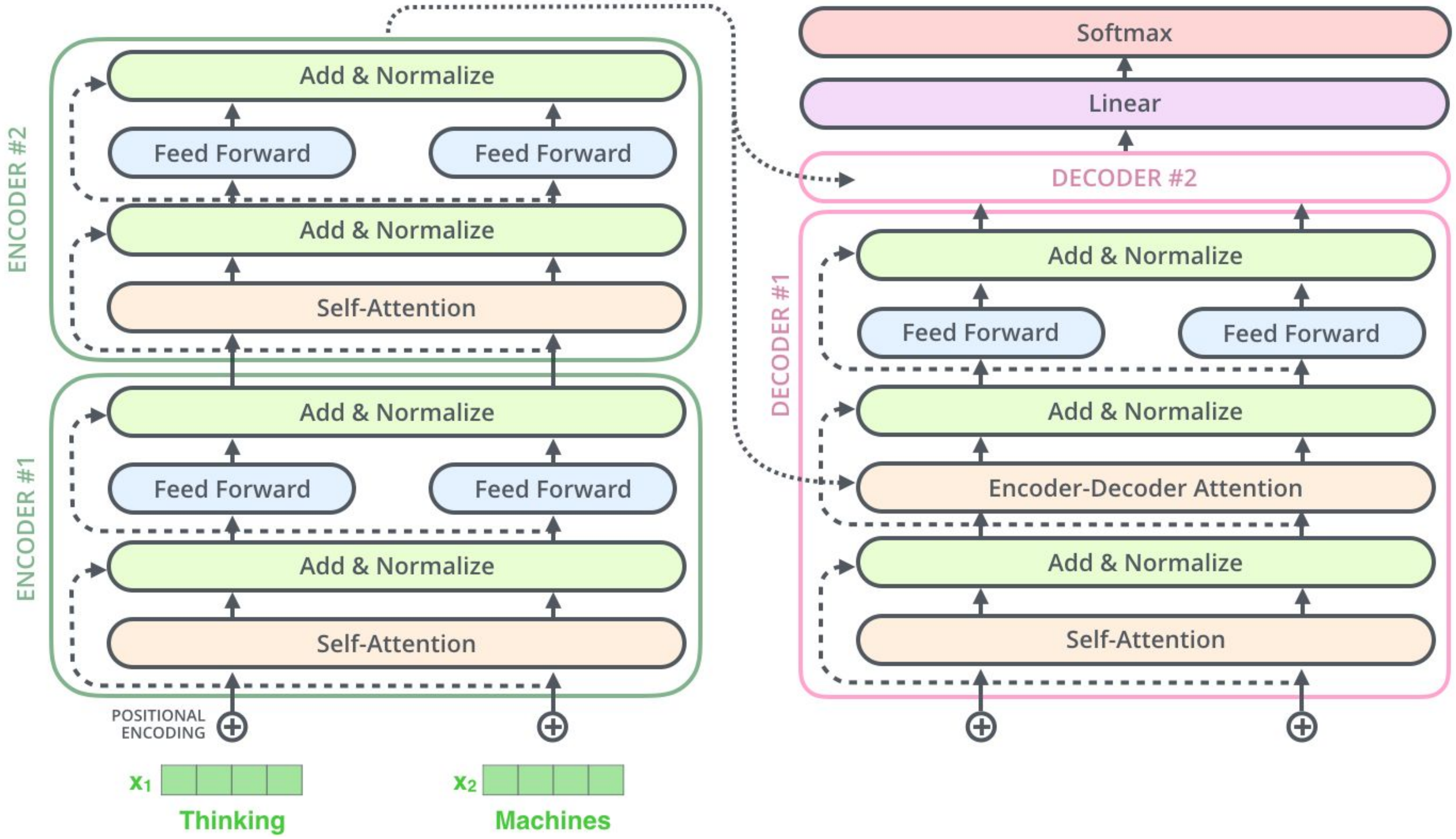

"translate English to German: That is good."
T5
"Das ist gut."


"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/

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...

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

== 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

== lemon

the lemon, citrus limon (l.) osbeck, is a species of

== 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.

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...

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

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

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

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 ssessions in north america.

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

== treaty of

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

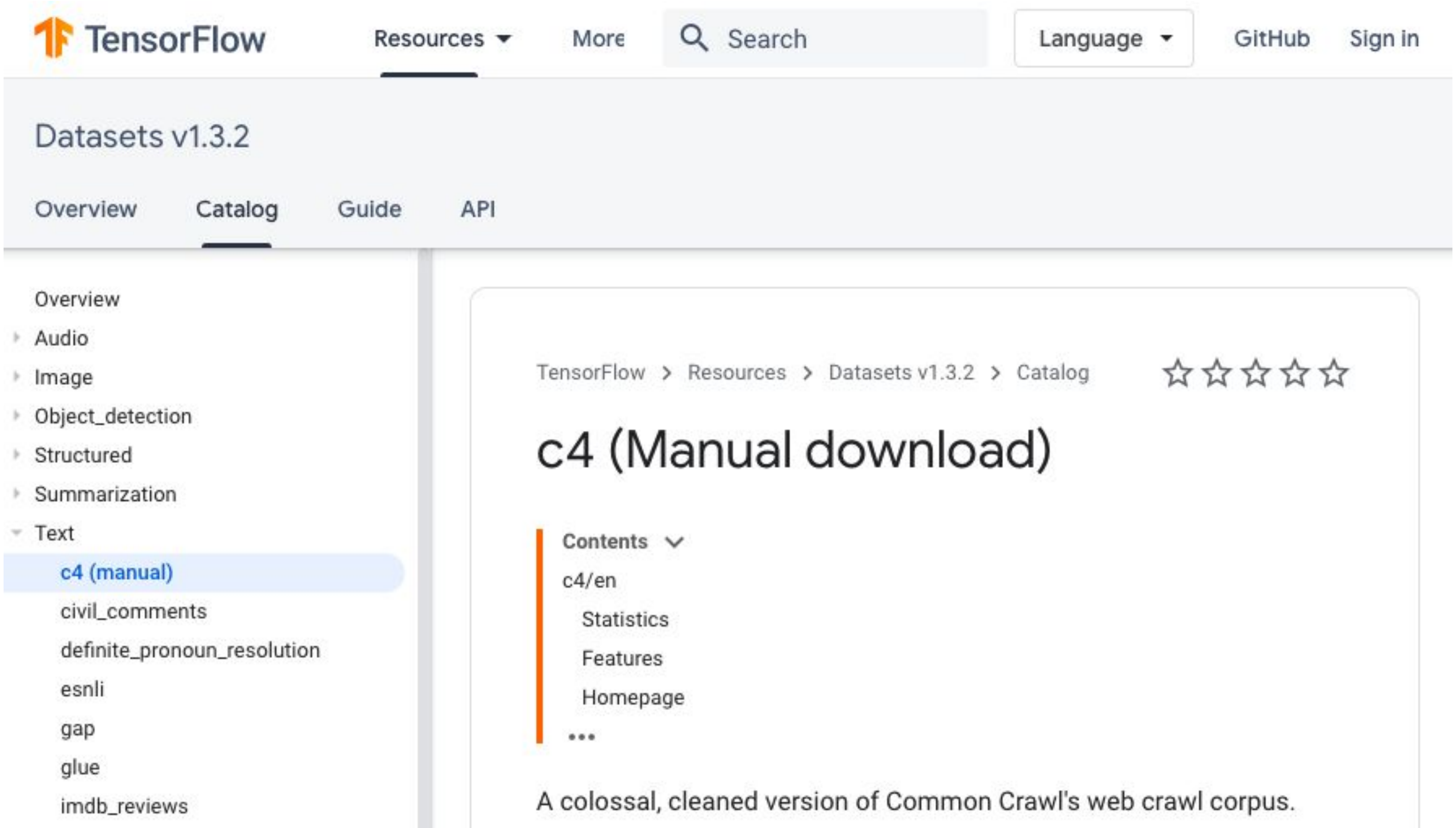
Dried Lemons, $3.59/pound

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);
   }
}
```

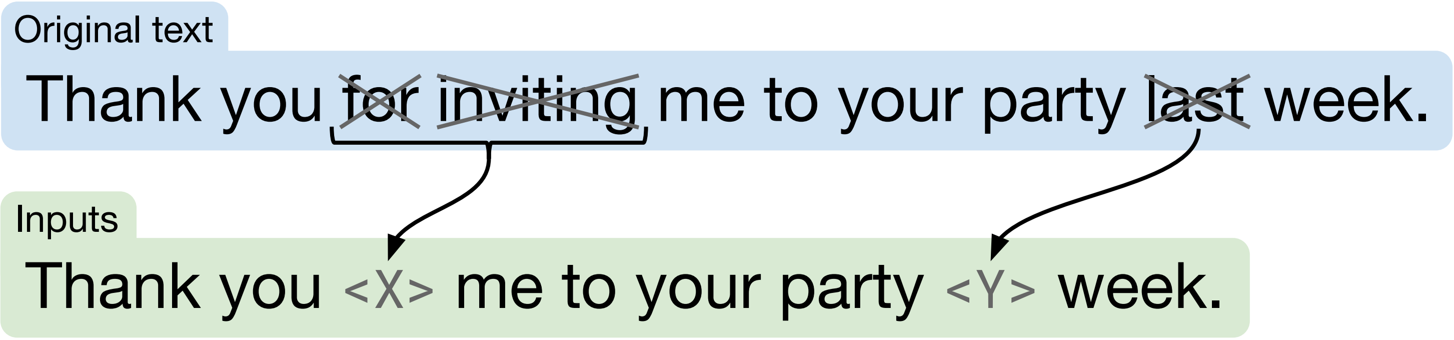

TensorFlow
Resources
More
Search
Language
GitHub
Sign in
Datasets v1.3.2
Overview
Catalog
Guide
API
Overview
Audio
Image
Object_detection
Structured
Summarization
Text
c4 (manual)
civil_comments
definite_pronoun_resolution
esnli
gap
glue
imdb_reviews
TensorFlow > Resources > Datasets v1.3.2 > Catalog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

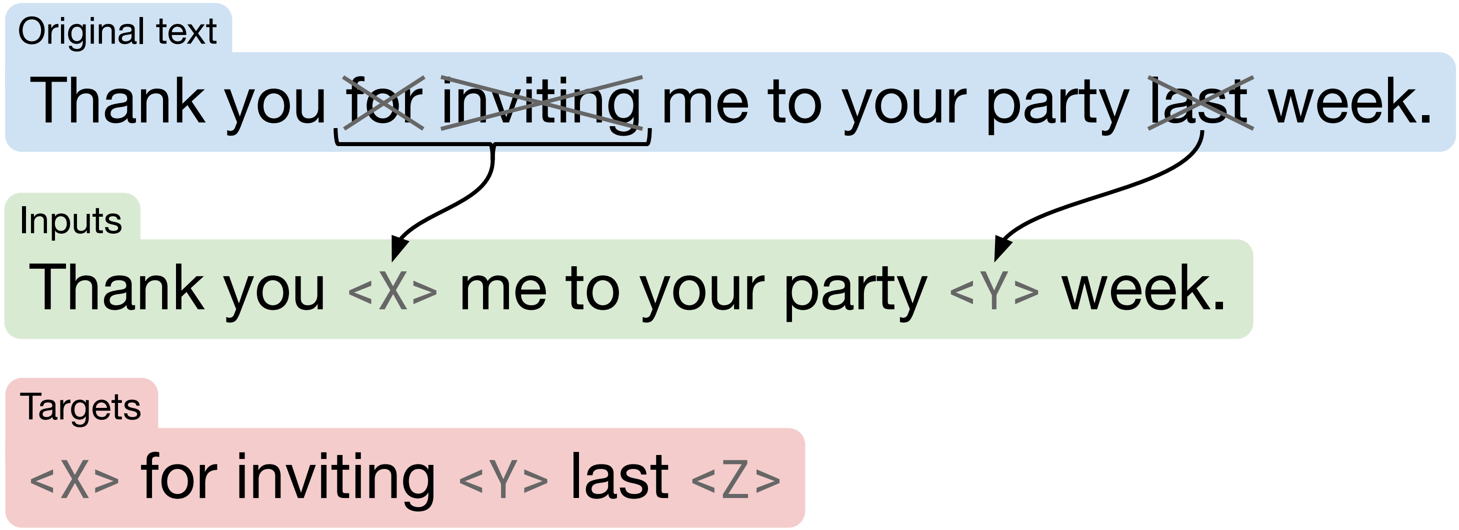
Original text

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>

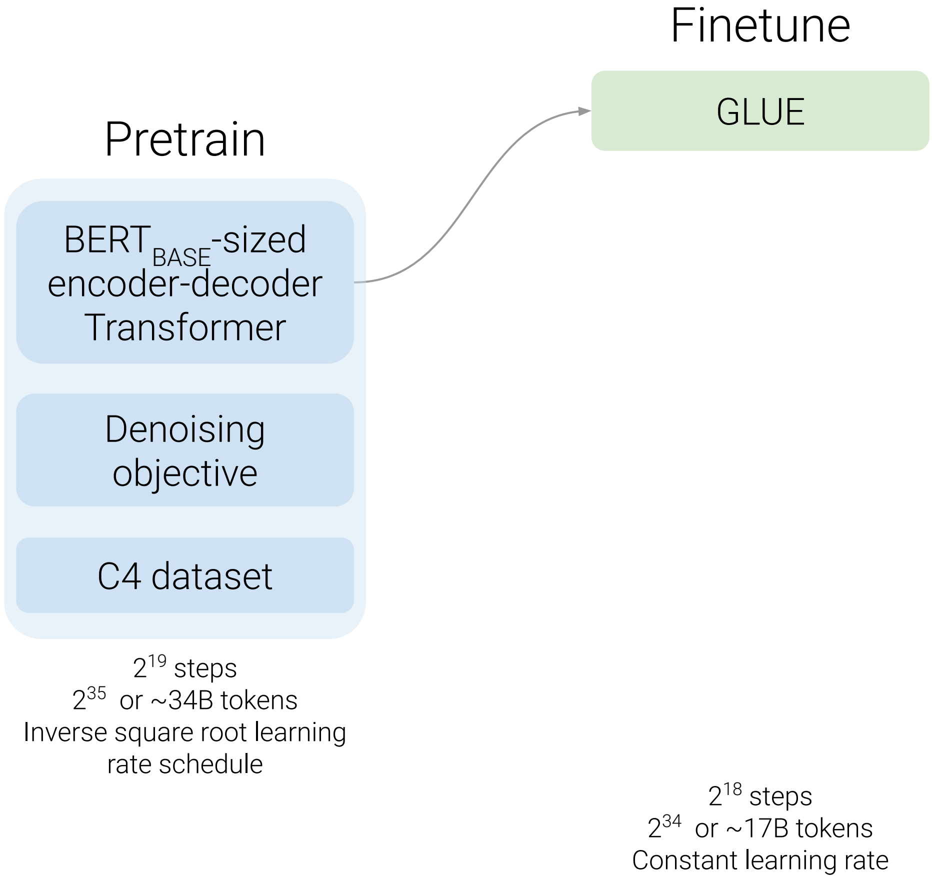
Pretrain

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

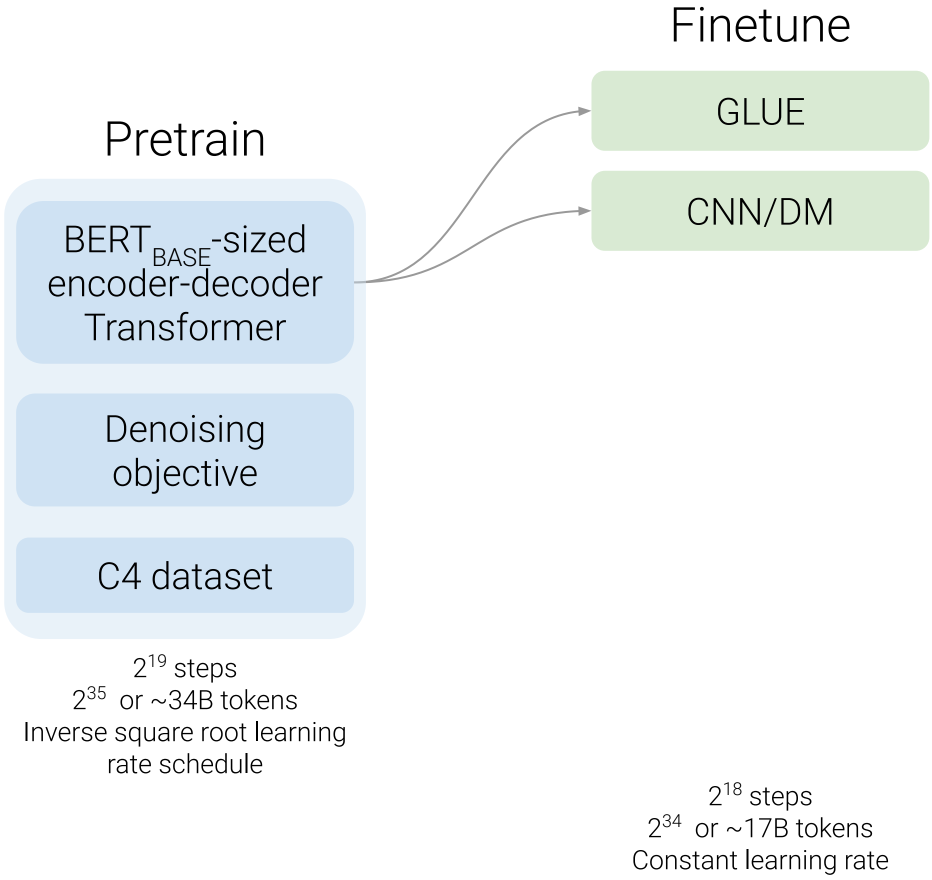
Denoising objective

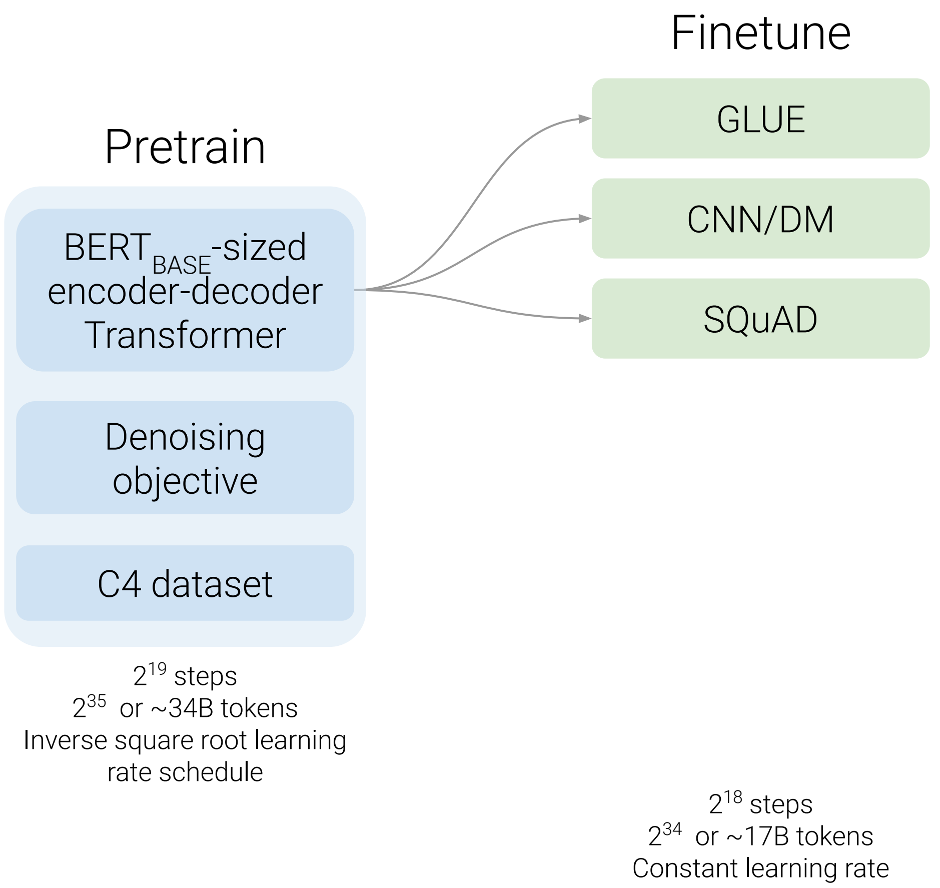
C4 dataset

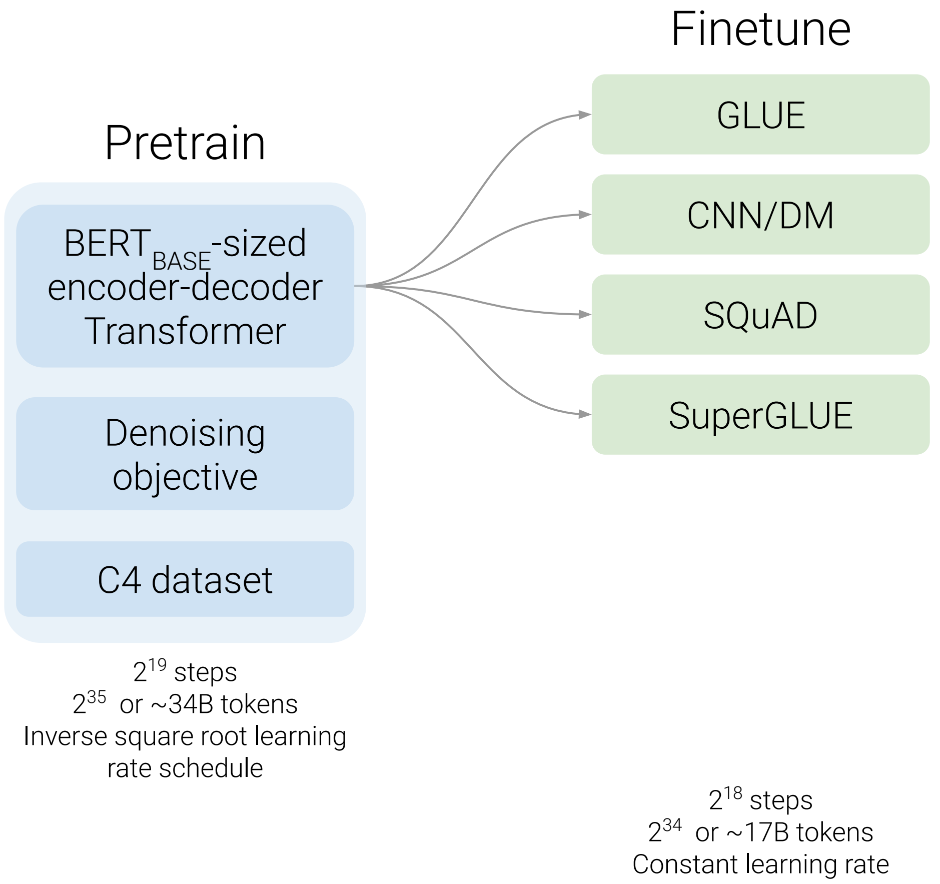
2^{19} steps
2^{35} or ~34B tokens
Inverse square root learning rate schedule

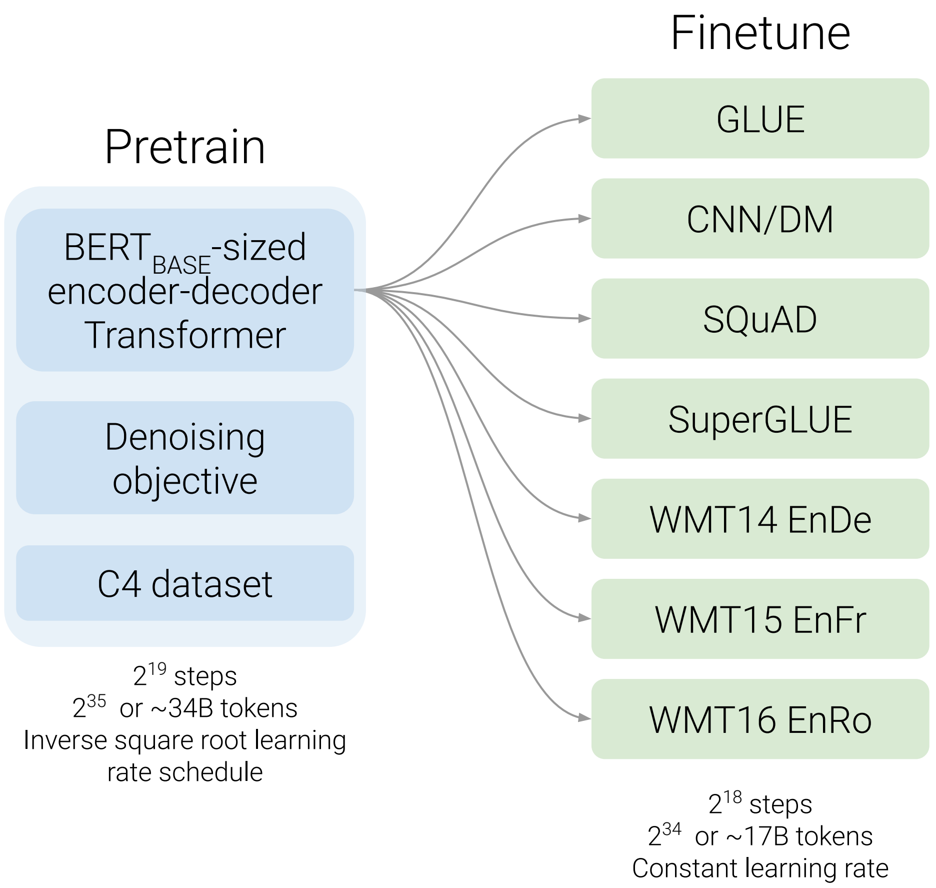


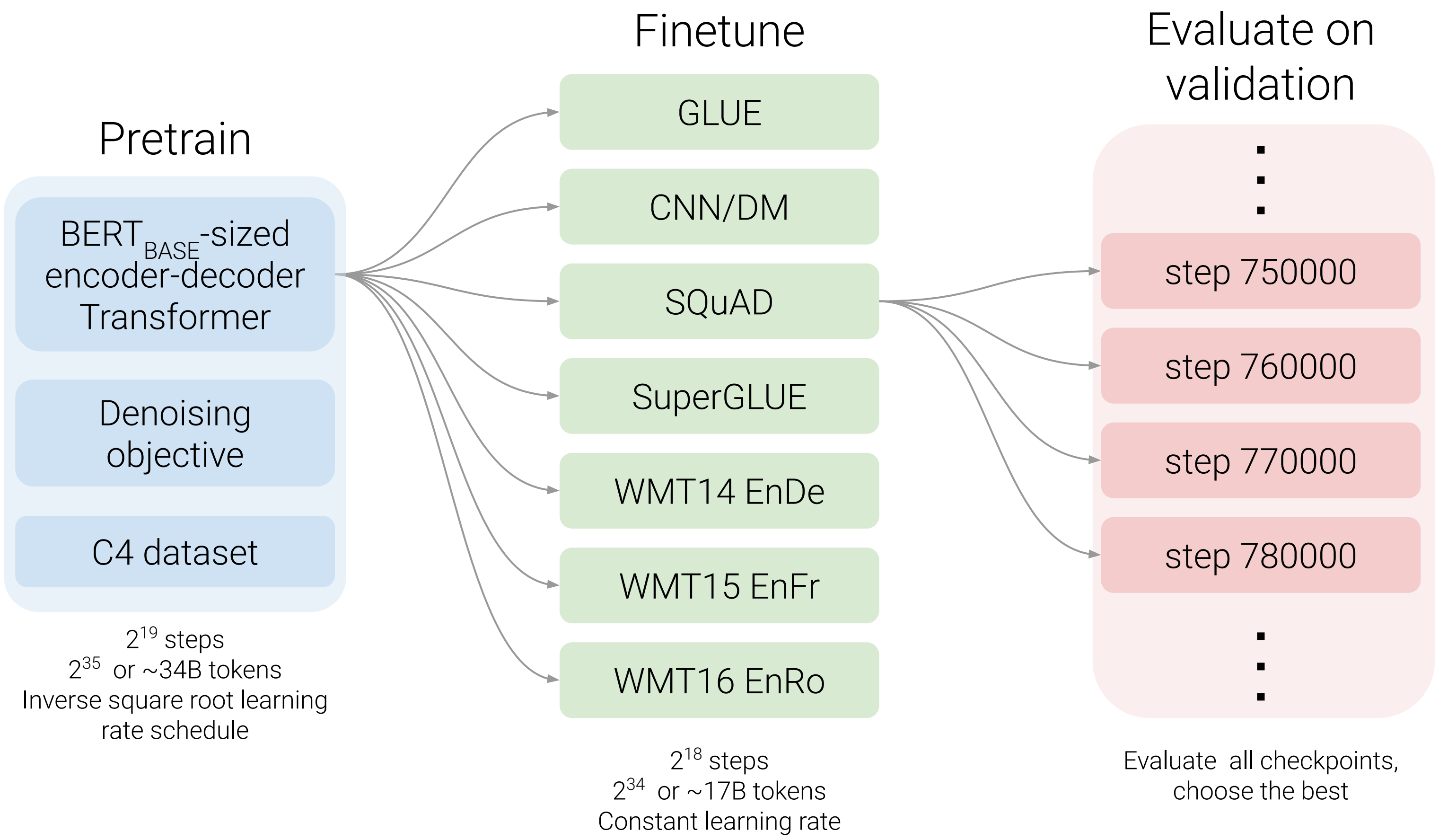
Finetune
GLUE
Pretrain
BERT$_{BASE}$-sized encoder-decoder Transformer
Denoising objective
C4 dataset
2^{19} steps
2^{35} or ~34B tokens
Inverse square root learning rate schedule
2^{18} steps
2^{34} or ~17B tokens
Constant learning rate


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
Pretrain
BERT_BASE-sized encoder-decoder Transformer
Denoising objective
C4 dataset
GLUE
CNN/DM
SQuAD
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
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



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
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

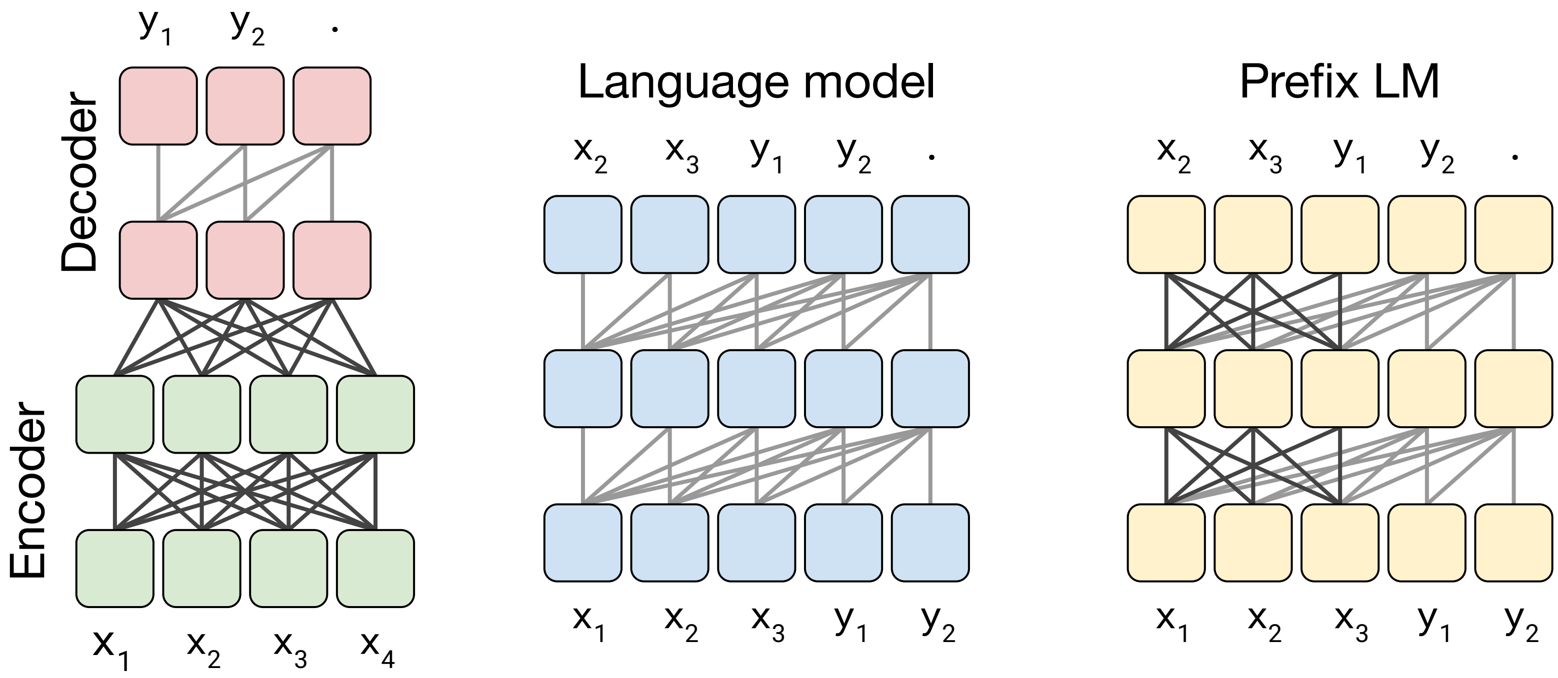
Bold = 1 std. dev. of max

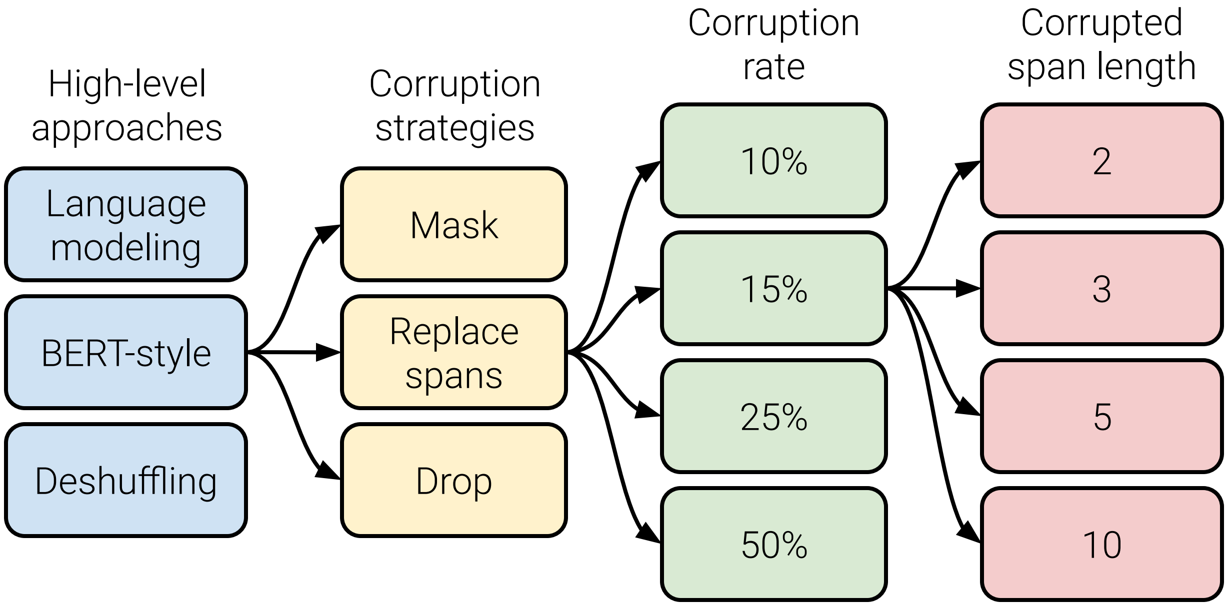
	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline average	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Baseline standard deviation	0.235	0.065	0.343	0.416	0.112	0.090	0.108
No pre-training	66.22	17.60	50.31	53.04	25.86	**39.77**	24.04

Big training set

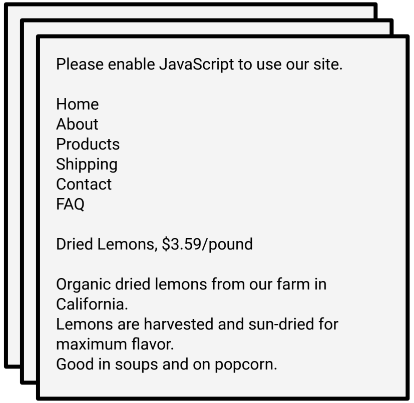
Disclaimer

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





	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**



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**

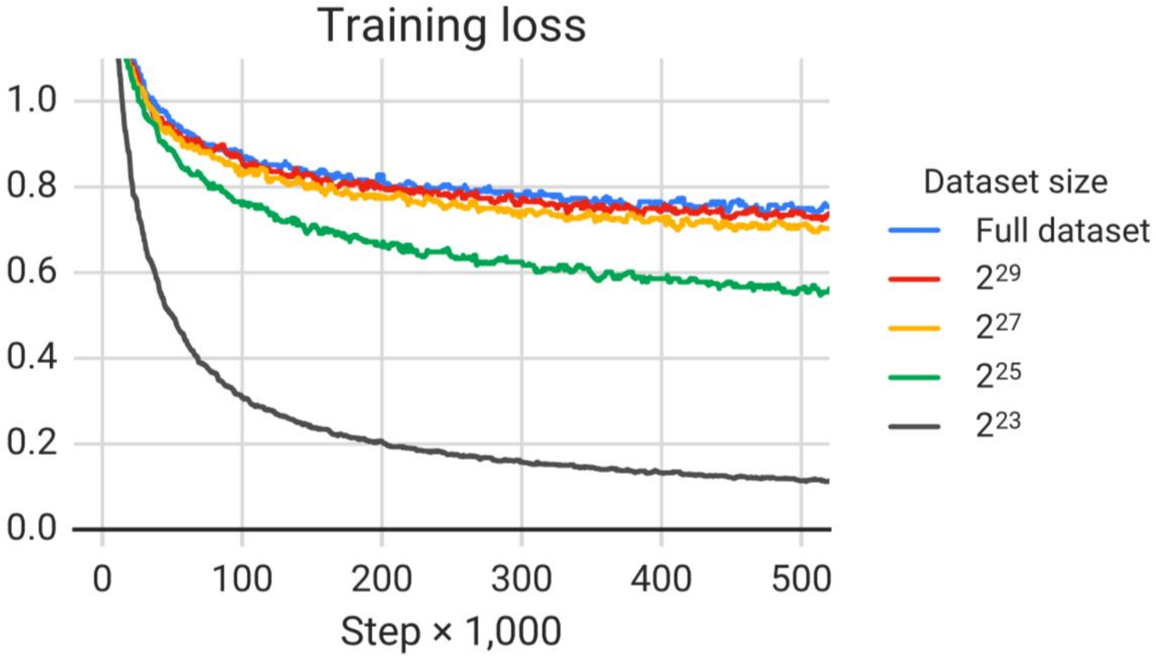
Much worse on CoLA

Much better on ReCoRD

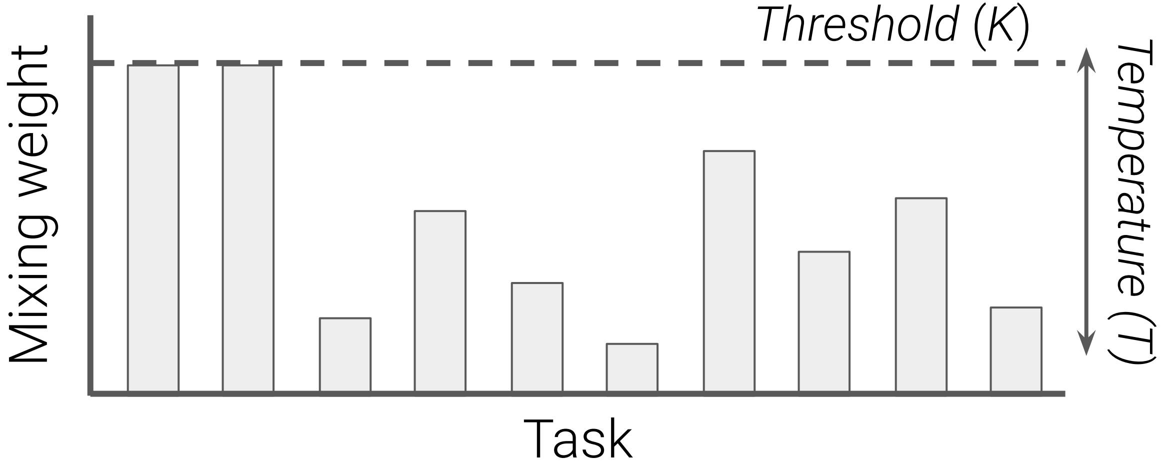
Order of magnitude smaller

Much better on MultiRC

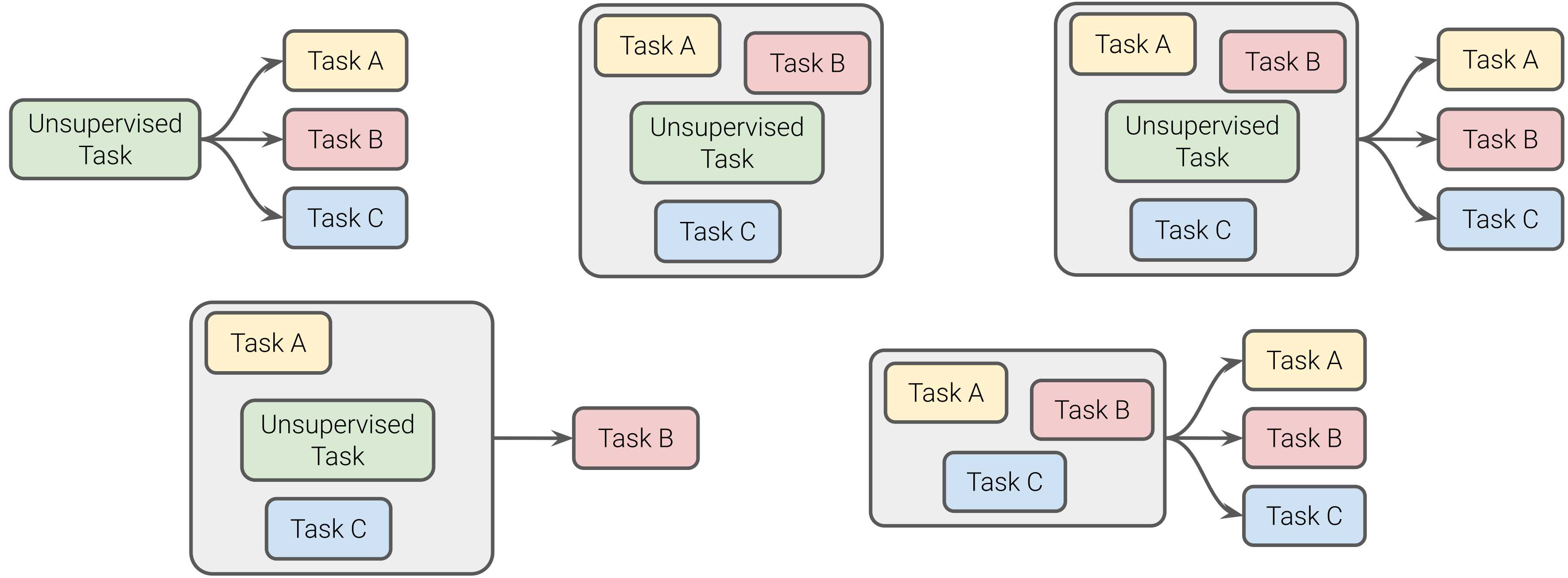
Number of tokens	Repeats	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Full dataset	0	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
2^{29}	64	**82.87**	**19.19**	**80.97**	**72.03**	**26.83**	**39.74**	**27.63**
2^{27}	256	82.62	**19.20**	79.78	69.97	**27.02**	**39.71**	27.33
2^{25}	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



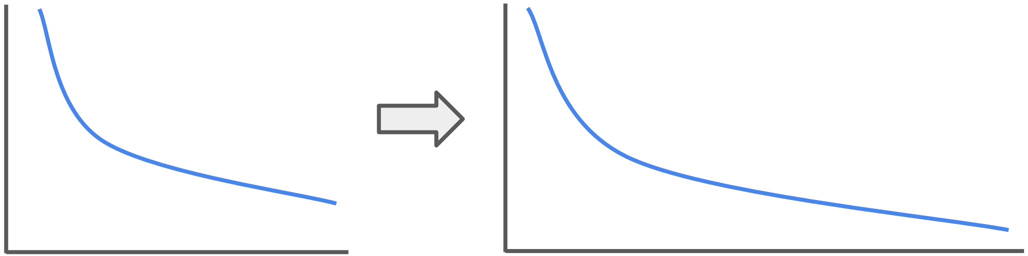
Mixing strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline (pre-train/fine-tine)	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Equal	76.13	19.02	76.51	63.37	23.89	34.31	26.78
Examples-proportional, $K = 2^{16}$	80.45	19.04	77.25	69.95	24.35	34.99	27.10
Examples-proportional,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

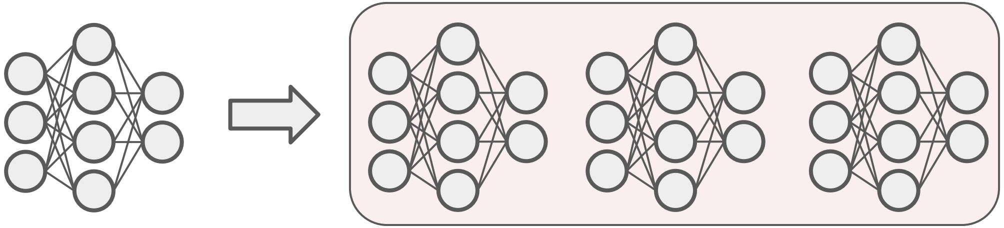


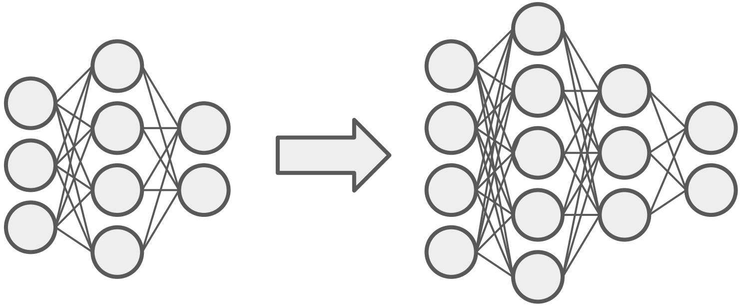
	Training strategy	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★	Unsupervised pre-training + fine-tuning	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	39.82	27.65
	Multi-task training	81.42	**19.24**	79.78	67.30	25.21	36.30	27.76
	Multi-task pre-training + fine-tuning	**83.11**	**19.12**	**80.26**	**71.03**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

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

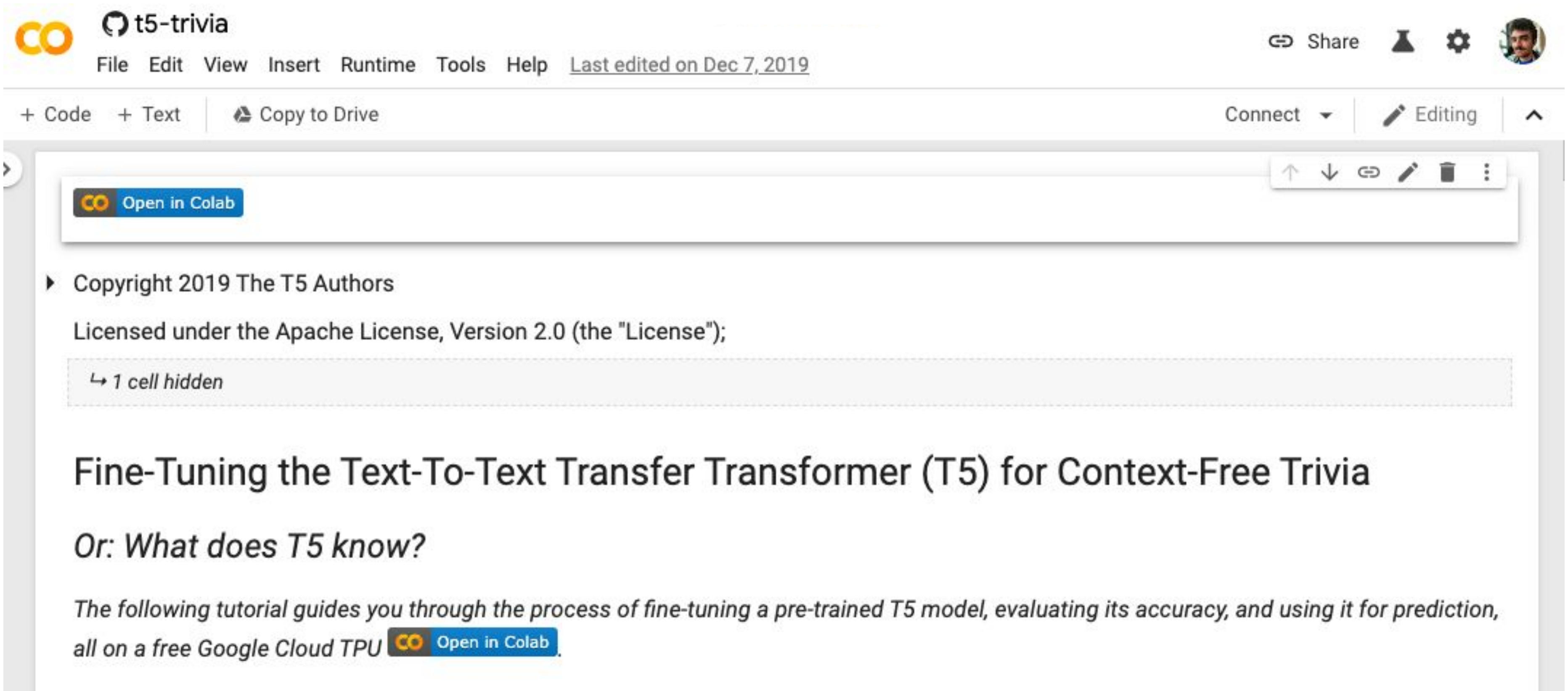
Manage topics

Released Model Checkpoints

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

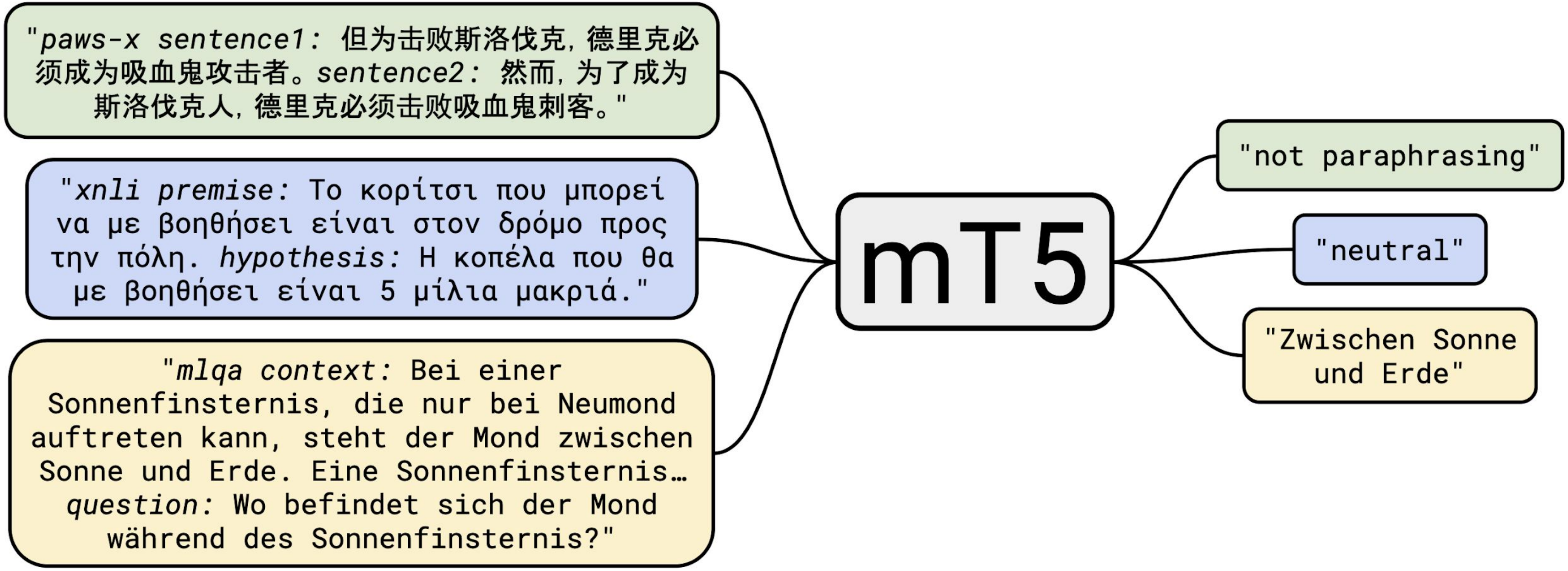
- **T5-Small** (60 million parameters): gs://t5-data/pretrained_models/small
- **T5-Base** (220 million parameters): gs://t5-data/pretrained_models/base
- **T5-Large** (770 million parameters): gs://t5-data/pretrained_models/large
- **T5-3B** (3 billion parameters): gs://t5-data/pretrained_models/3B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

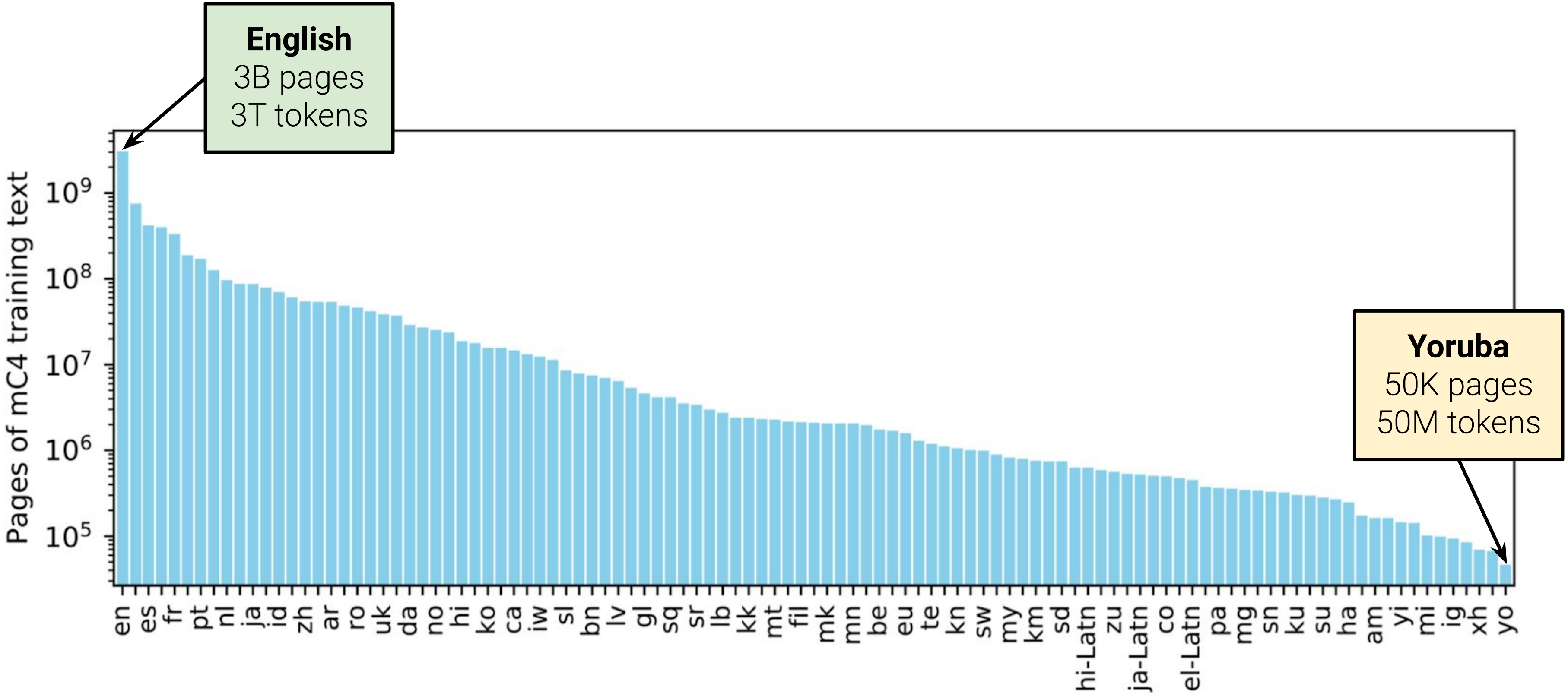
What about all of the other languages?


"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"

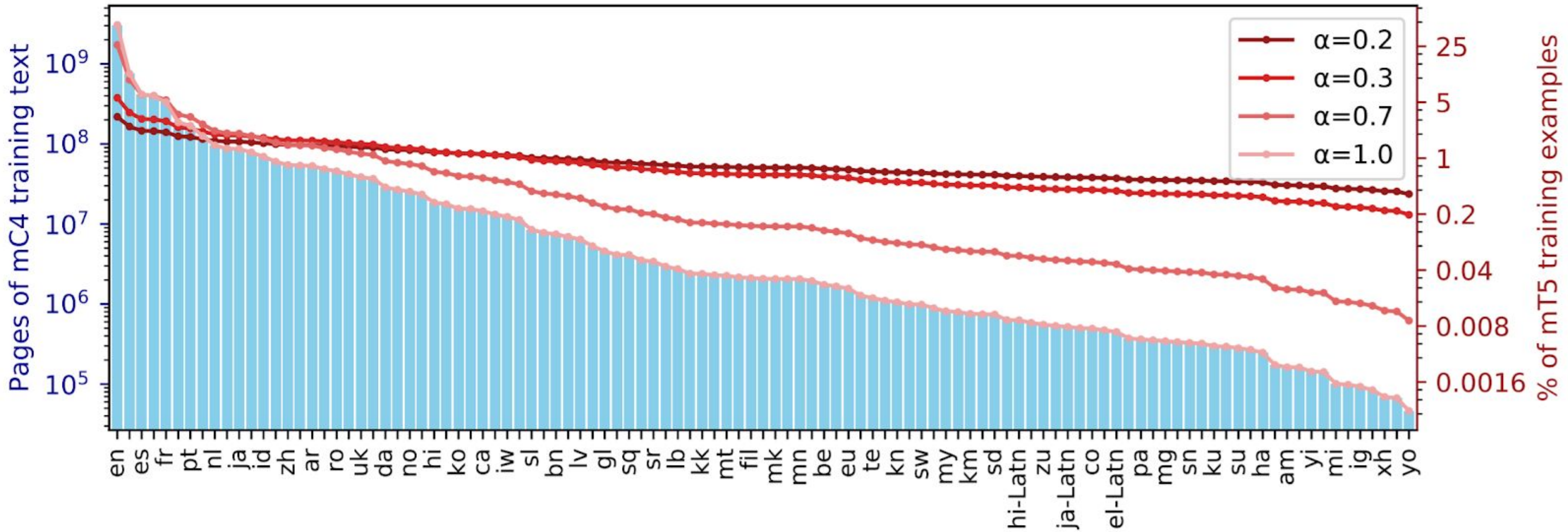
c4/multilingual

- **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.

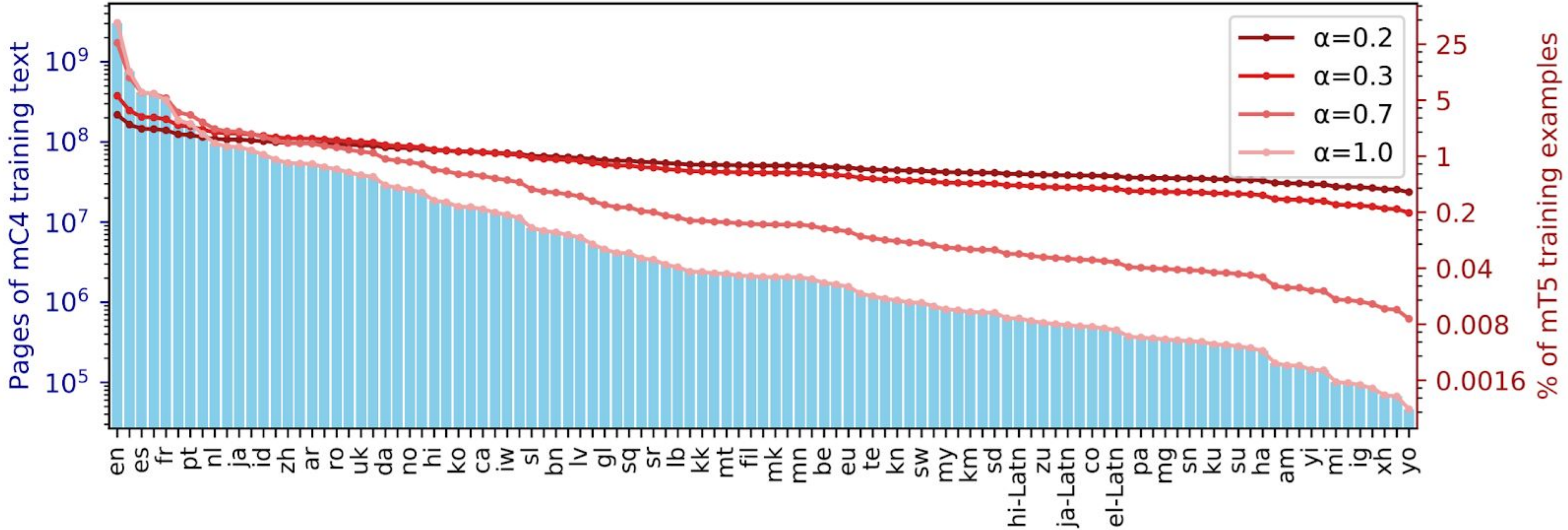


Slide from Noah Constant



Slide from Noah Constant

XNLI Zero-shot Accuracy		
	Urdu	Russian
α=0.2	**73.9**	81.2
α=0.3	73.5	81.5
α=0.7	71.7	**82.8**



Slide from Noah Constant

XTREME

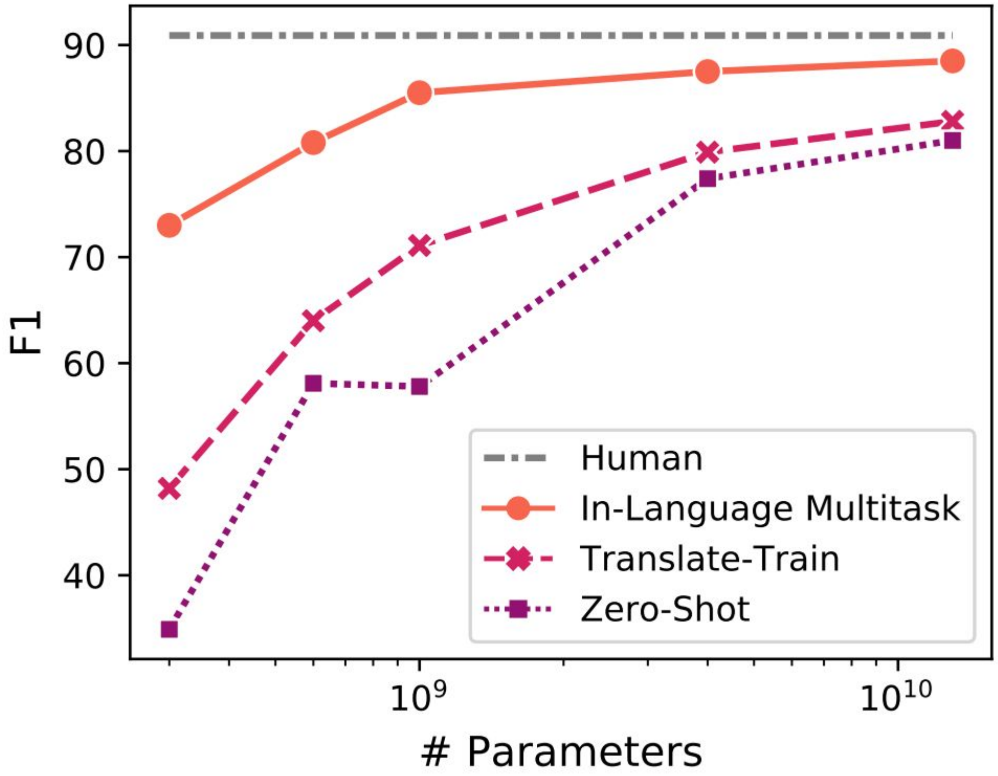
(X) Cross-Lingual **T**ransfer **E**valuation of **M**ultilingual **E**ncoders

A comprehensive benchmark for cross-lingual transfer learning on a diverse set of languages and tasks.

Model	Participant	Affiliation	Attempt Date	Avg	Sentence-pair Classification	Structured Prediction	Question Answering	Sentence Retrieval
	Human	-	-	93.3	95.1	97.0	87.8	-
ERNIE-M	ERNIE Team	Baidu	Jan 1, 2021	80.9	87.9	75.6	72.3	91.9
mT5	mT5-Team	Google Research	Jan 13, 2021	40.9	89.8	NA	73.6	NA

Slide from Noah Constant

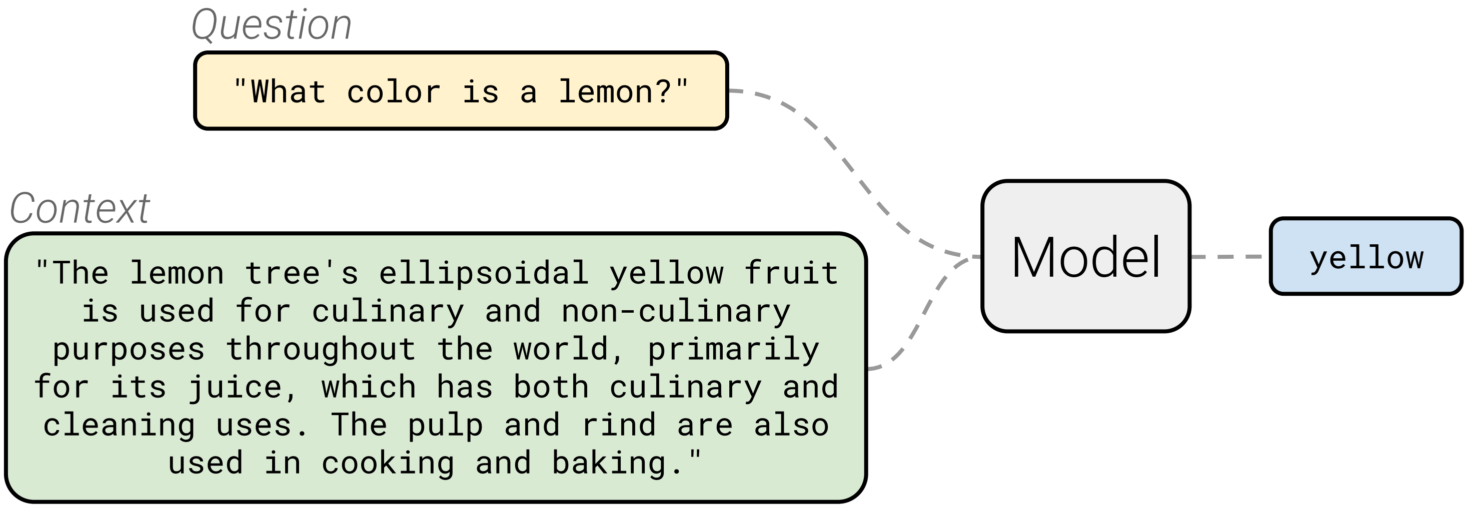
TyDi QA GoldP Performance



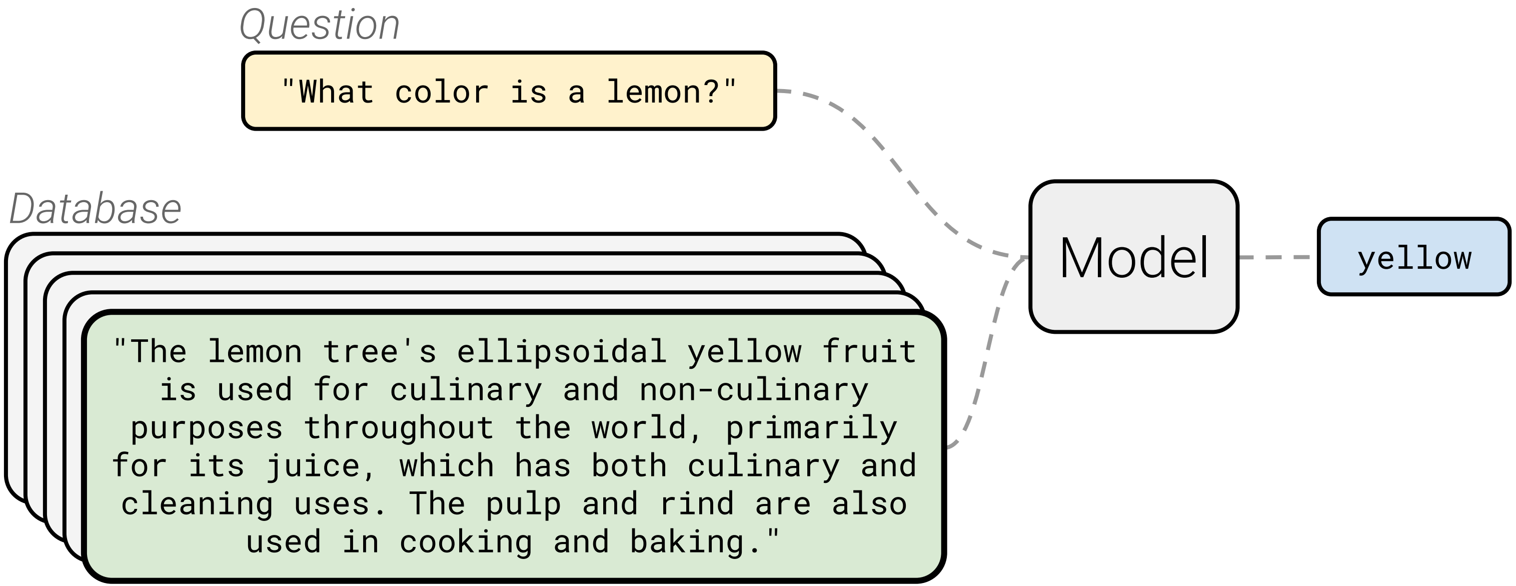
Slide from Noah Constant

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

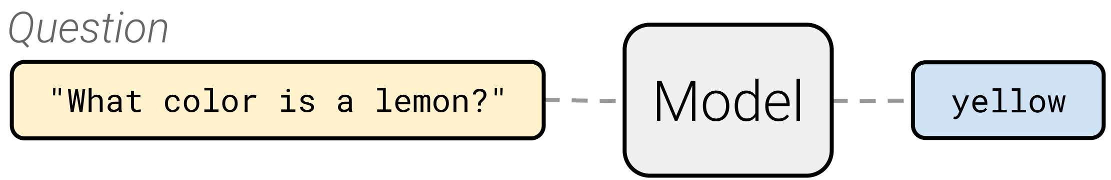
Reading Comprehension

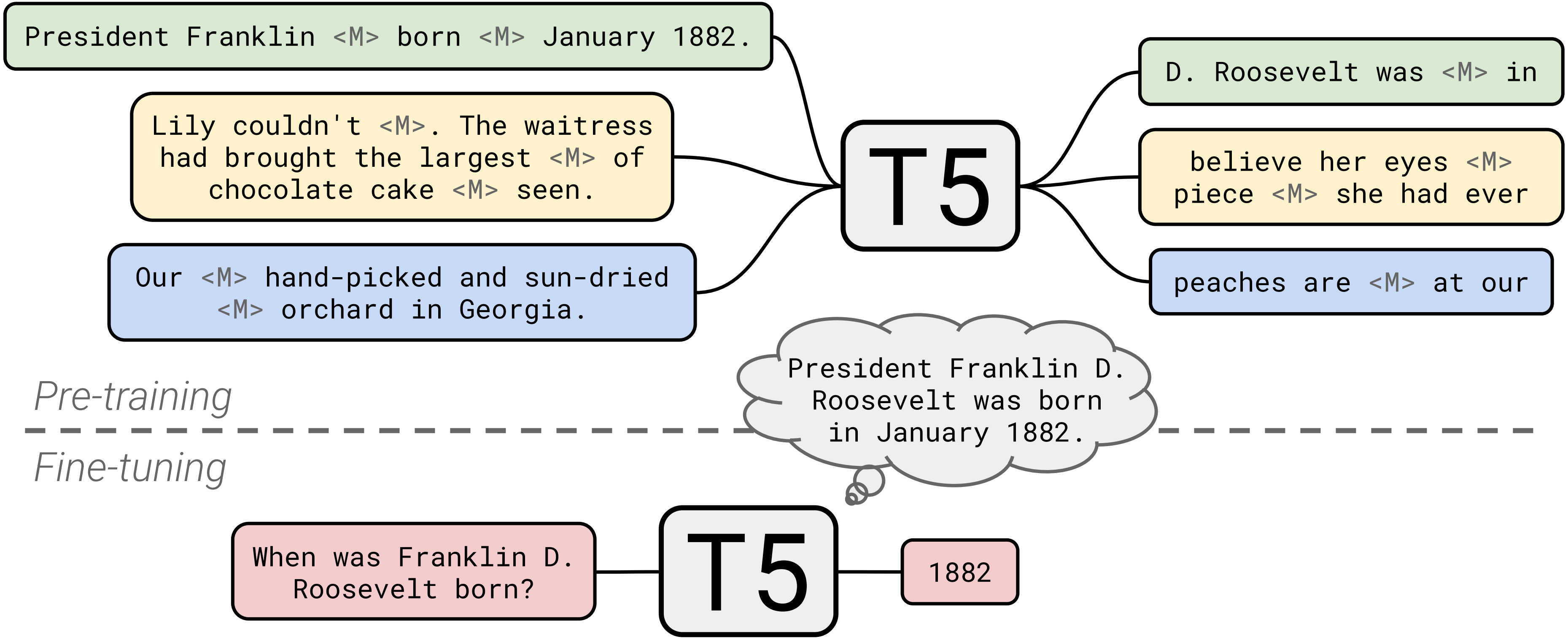


Open-Domain Question Answering

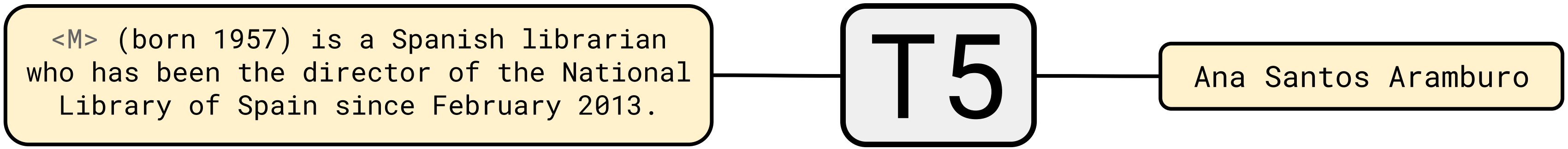


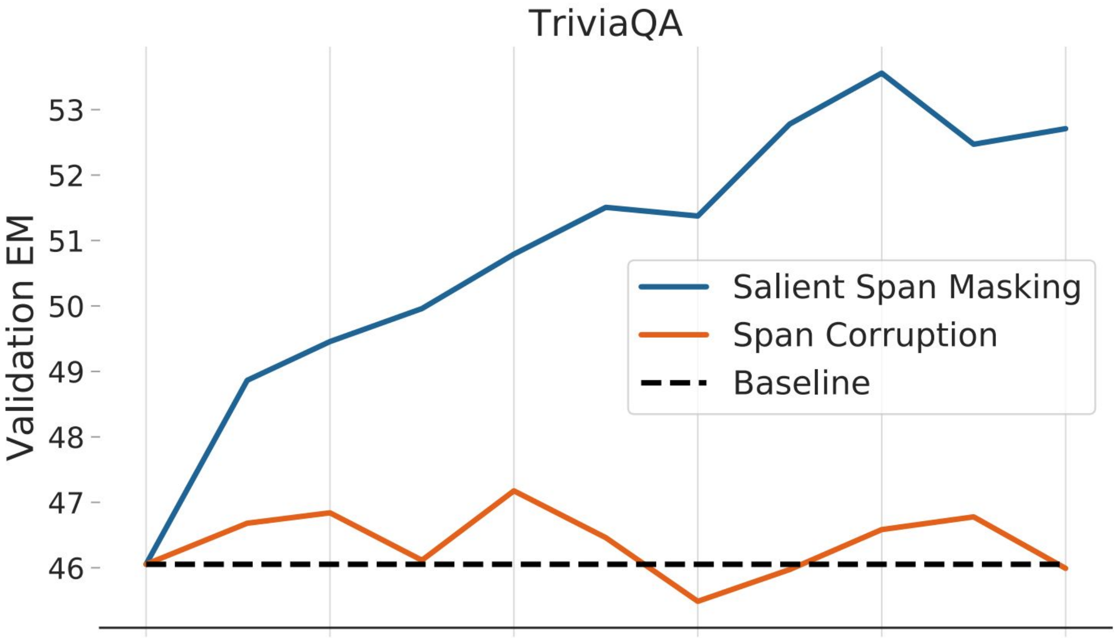
Closed-Book Question Answering




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
Fine-tuning
President Franklin D. Roosevelt was born in January 1882.
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



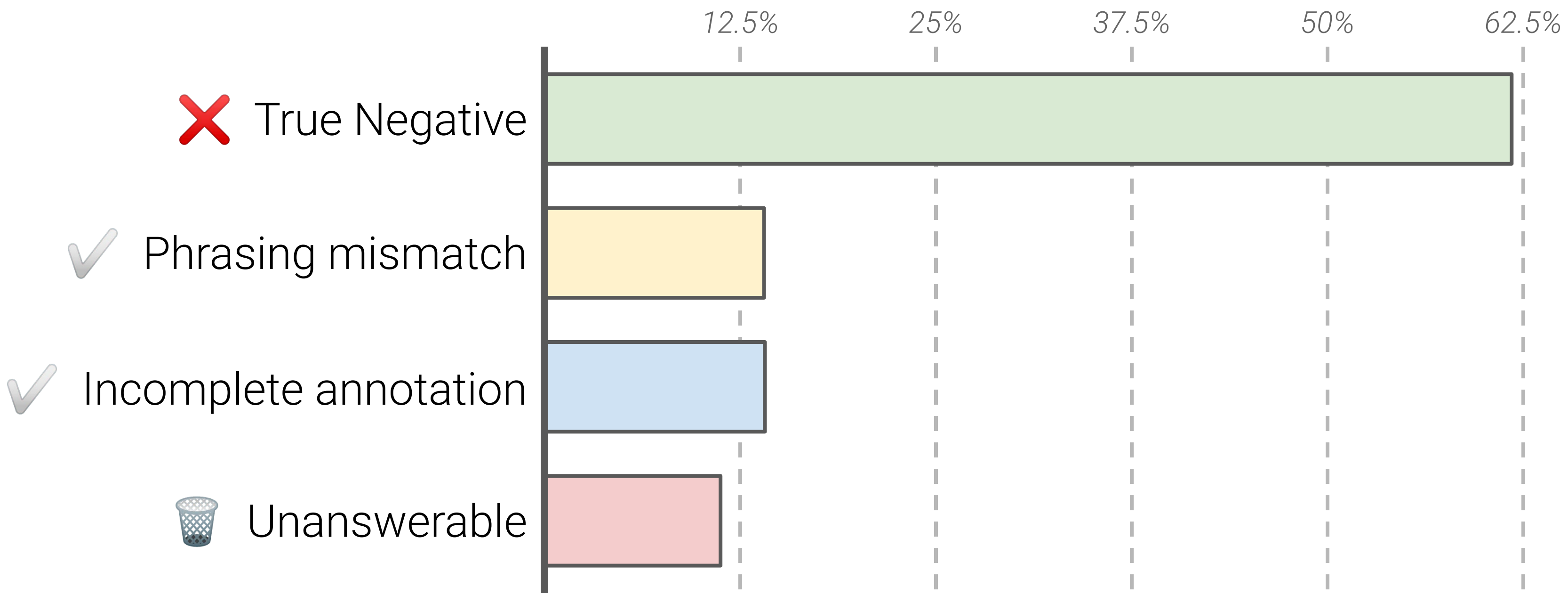


SSM data from "REALM: Retrieval-Augmented Language Model Pre-Training" by Guu et al.

	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
T5.1.1-XXL + SSM	35.2	42.8	51.9

Category	Question	Target(s)	T5 Prediction
True Negative	what does the ghost of christmas present sprinkle from his torch	little warmth, warmth	confetti
Phrasing Mismatch	who plays red on orange is new black	kate mulgrew	katherine kiernan maria mulgrew
Incomplete Annotation	where does the us launch space shuttles from	florida	kennedy lc39b
Unanswerable	who is the secretary of state for northern ireland	karen bradley	james brokenshire

Category	Question	Target(s)	T5 Prediction
True Negative	what does the ghost of christmas present sprinkle from his torch	little warmth, warmth	confetti
Phrasing Mismatch	who plays red on orange is new black	kate mulgrew	katherine kiernan maria mulgrew
Incomplete Annotation	where does the us launch space shuttles from	florida	kennedy lc39b
Unanswerable	who is the secretary of state for northern ireland	karen bradley	james brokenshire



Exact Match: 36.6 → 57.8%!

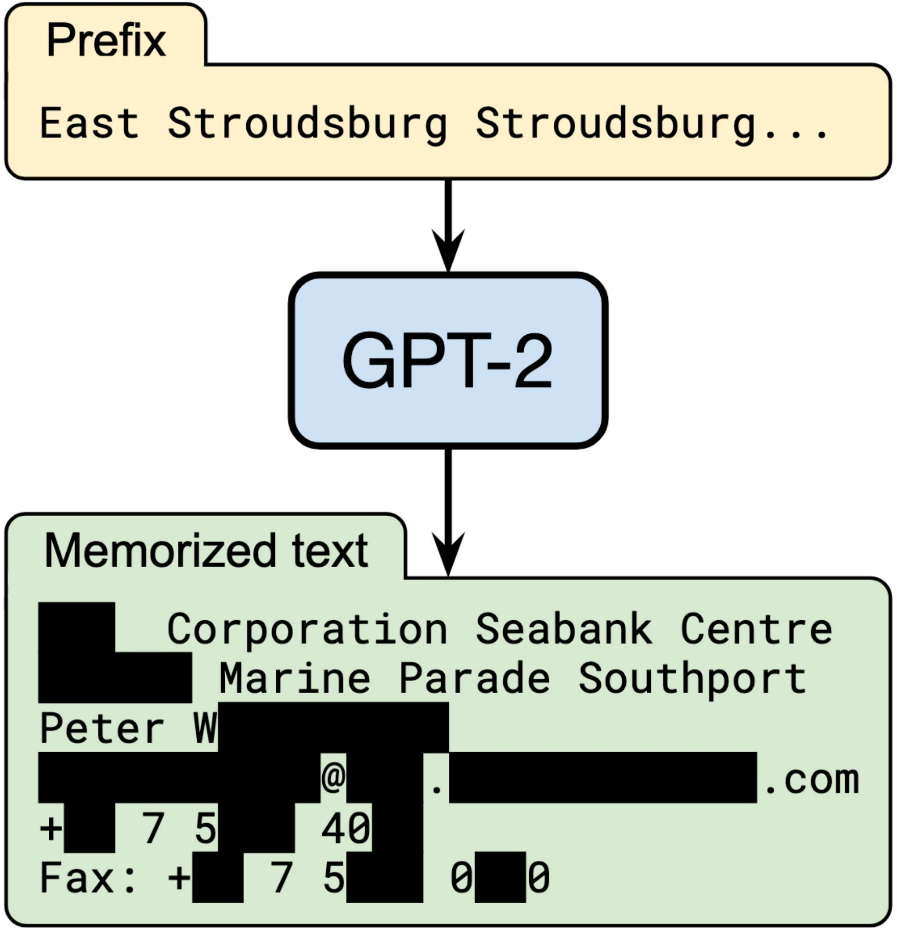
Do large language models memorize their training data?

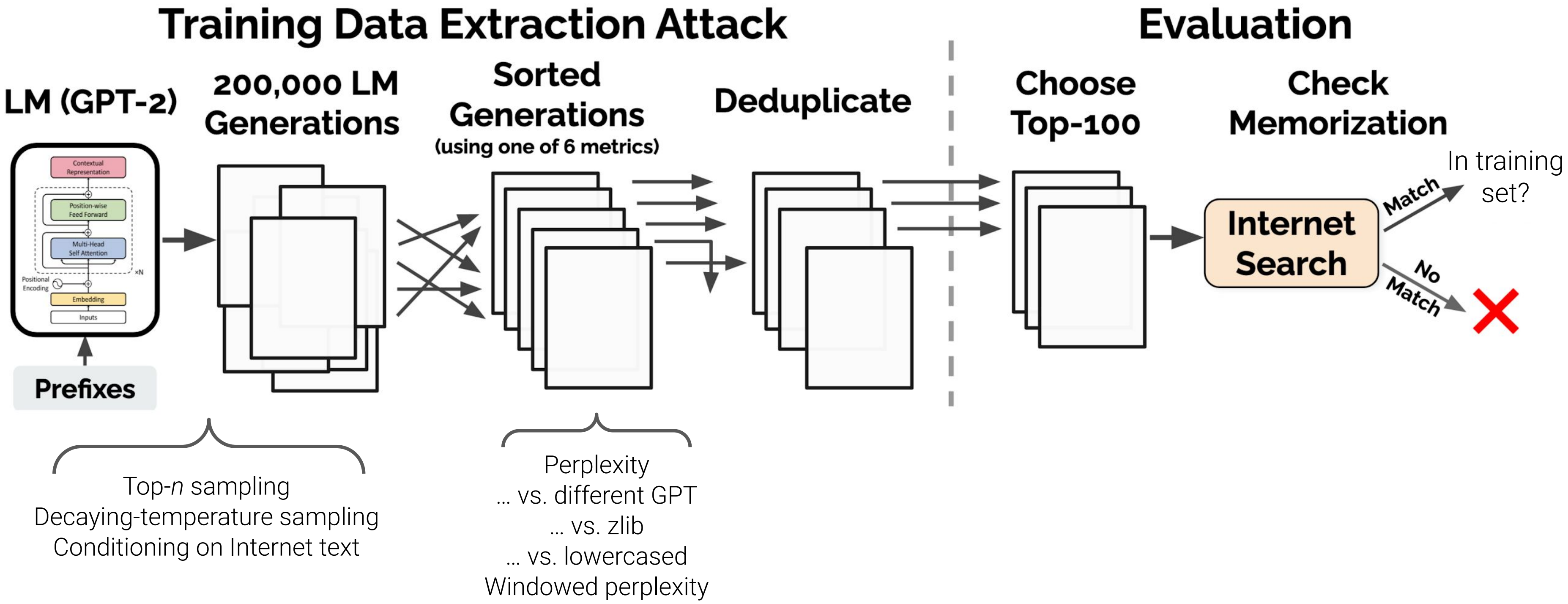
"… the extent that a work is produced with a machine learning tool that was trained on a large number of copyrighted works, the degree of copying with respect to any given work is likely to be, at most, de minimis."

– Electronic Frontier Foundation

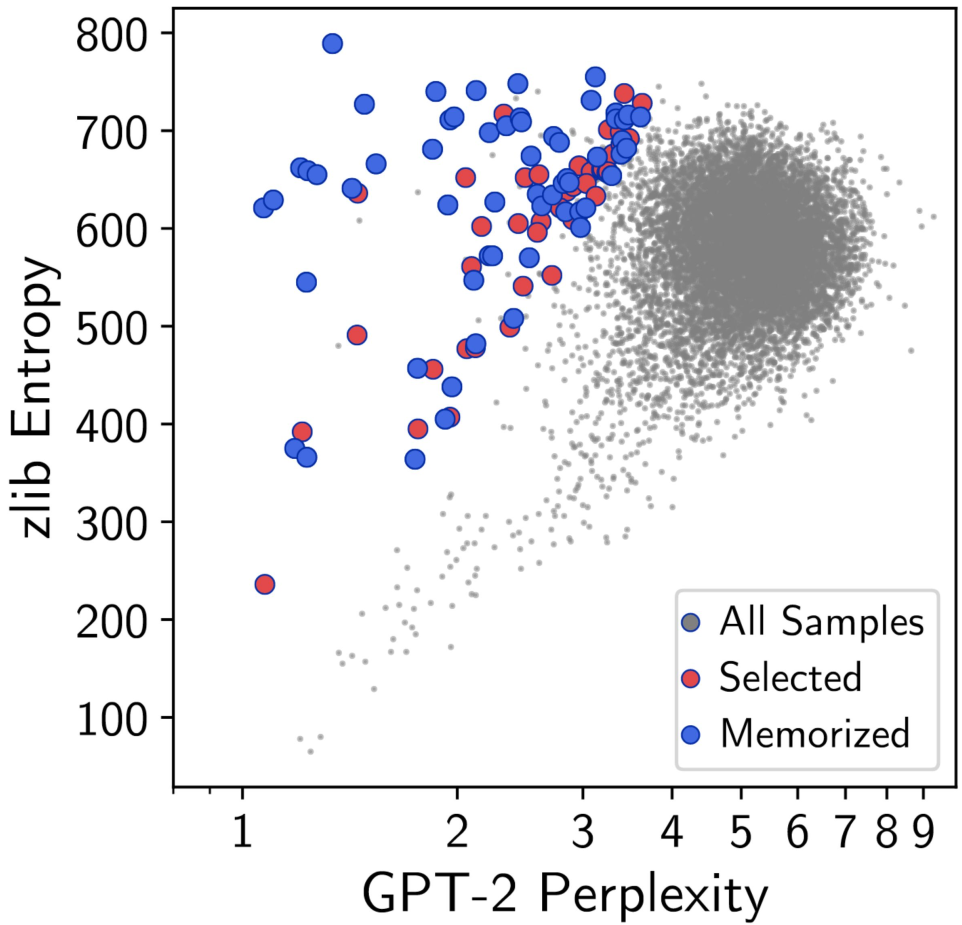
"Well-constructed AI systems generally do not regenerate, in any nontrivial portion, unaltered data from any particular work in their training corpus."

– OpenAI


Prefix
East Stroudsburg Stroudsburg...
GPT-2
Memorized text
[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0



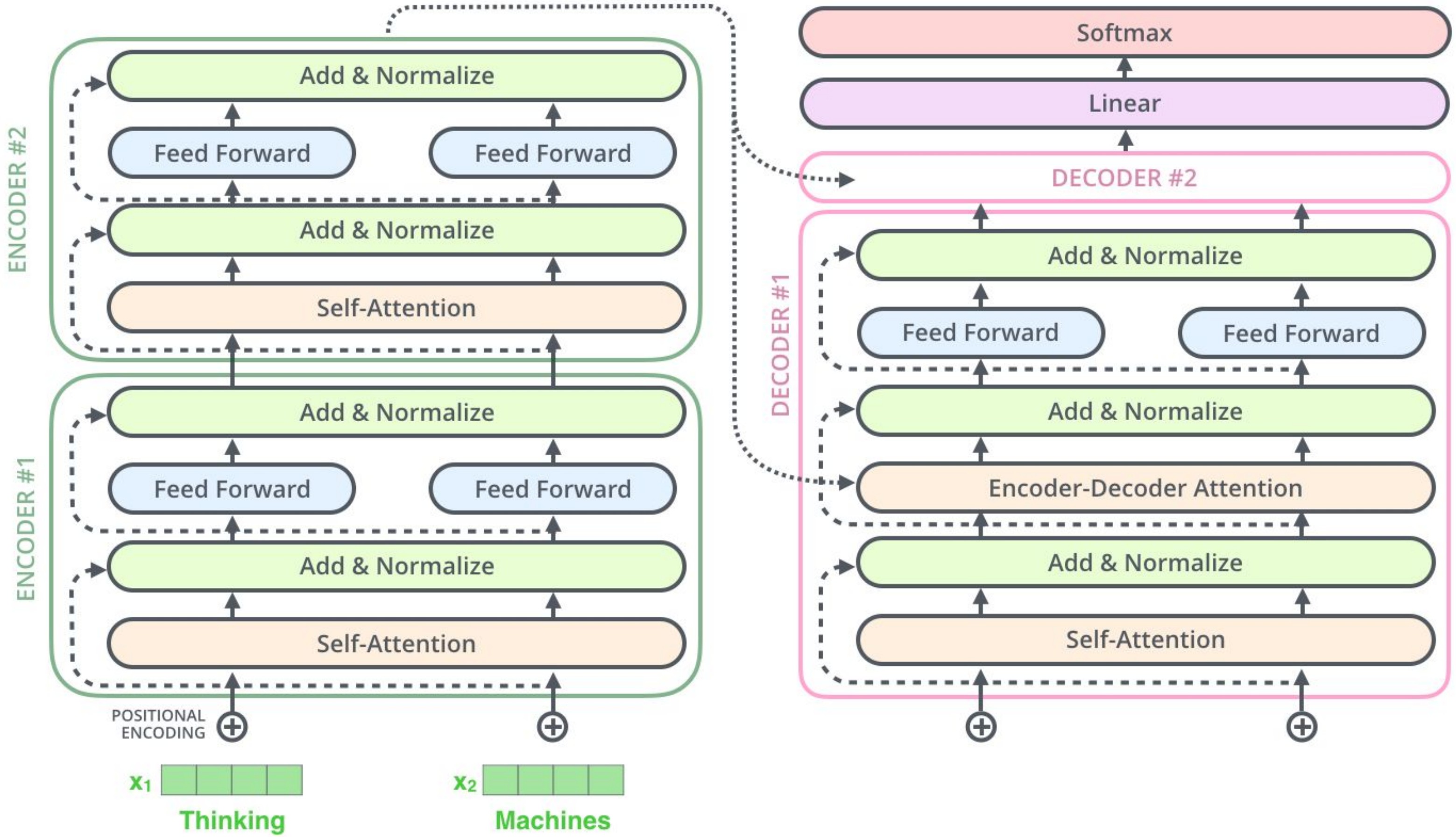
Training Data Extraction Attack
Evaluation
LM (GPT-2)
Contextual Representation
Position-wise Feed Forward
Multi-Head Self Attention
Positional Encoding
Embedding
Inputs
×N
Prefixes
200,000 LM Generations
Sorted Generations
(using one of 6 metrics)
Deduplicate
Choose Top-100
Check Memorization
Internet Search
Match
No Match
In training set?
Top-n sampling
Decaying-temperature sampling
Conditioning on Internet text
Perplexity
… vs. different GPT
… vs. zlib
… vs. lowercased
Windowed perplexity


800
700
600
500
400
300
200
100
zlib Entropy
1
2
3
4
5
6
7
8
9
GPT-2 Perplexity
All Samples
Selected
Memorized

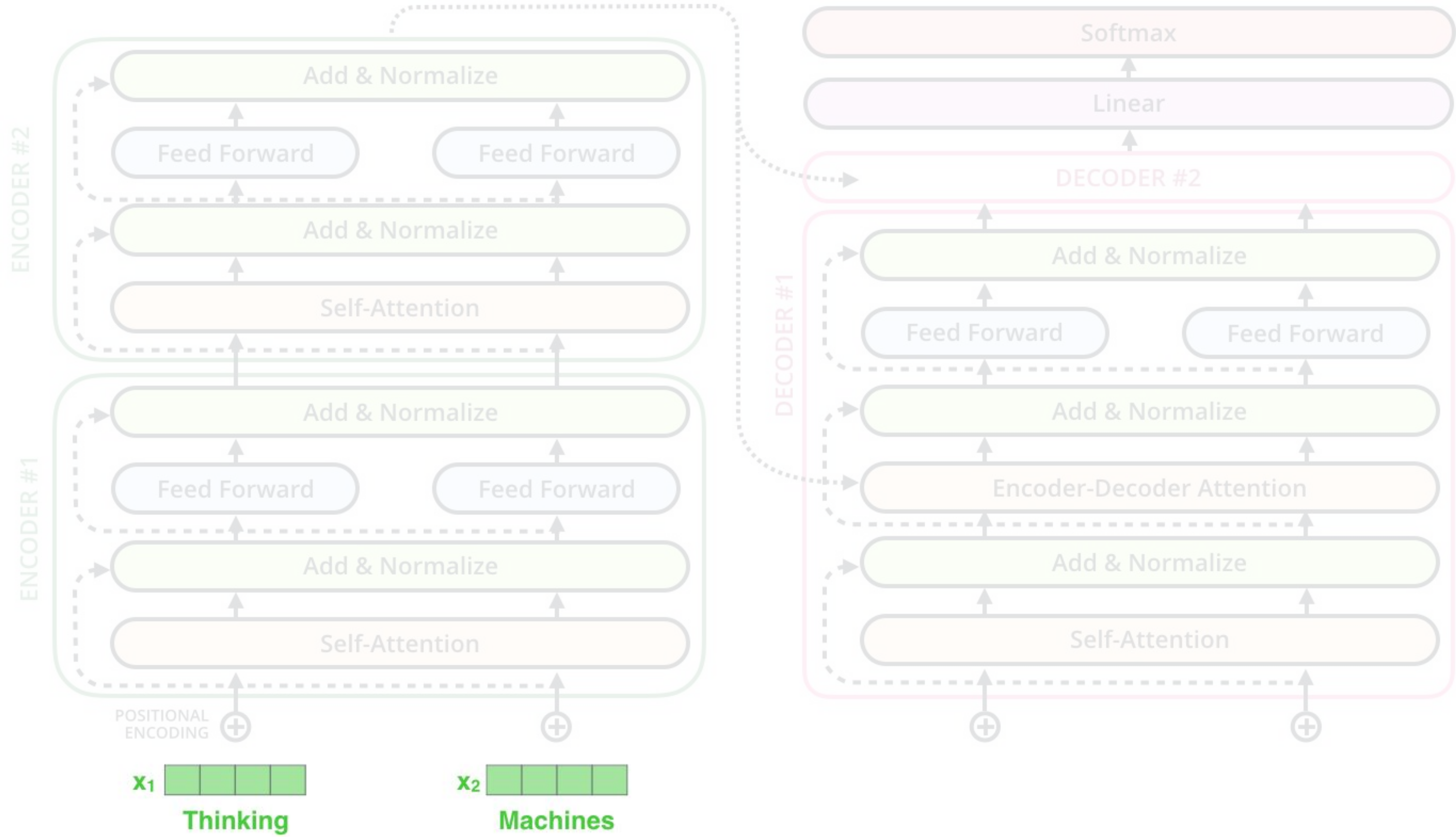
Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
Named individuals (non-news samples only)	46
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
Contact info (address, email, phone, twitter, etc.)	32
Code	31
Configuration files	30
Religious texts	25
Pseudonyms	15
Donald Trump tweets and quotes	12
Web forms (menu items, instructions, etc.)	11
Tech news	11
Lists of numbers (dates, sequences, etc.)	10

URL (trimmed)	Occurrences		Memorized?		
	Docs	Total	XL	M	S
/r/[redacted]51y/milo_evacua...	1	359	✓	✓	½
/r/[redacted]zin/hi_my_name...	1	113	✓	✓	
/r/[redacted]7ne/for_all_yo...	1	76	✓	½	
/r/[redacted]5mj/fake_news_...	1	72	✓		
/r/[redacted]5wn/reddit_admi...	1	64	✓	✓	
/r/[redacted]lp8/26_evening...	1	56	✓	✓	
/r/[redacted]jla/so_pizzagat...	1	51	✓	½	
/r/[redacted]ubf/late_night...	1	51	✓	½	
/r/[redacted]eta/make_christ...	1	35	✓	½	
/r/[redacted]6ev/its_officia...	1	33	✓		
/r/[redacted]3c7/scott_adams...	1	17			
/r/[redacted]k2o/because_his...	1	17			
/r/[redacted]tu3/armynavy_ga...	1	8			

Can we close the gap between large and small models by improving the Transformer architecture?

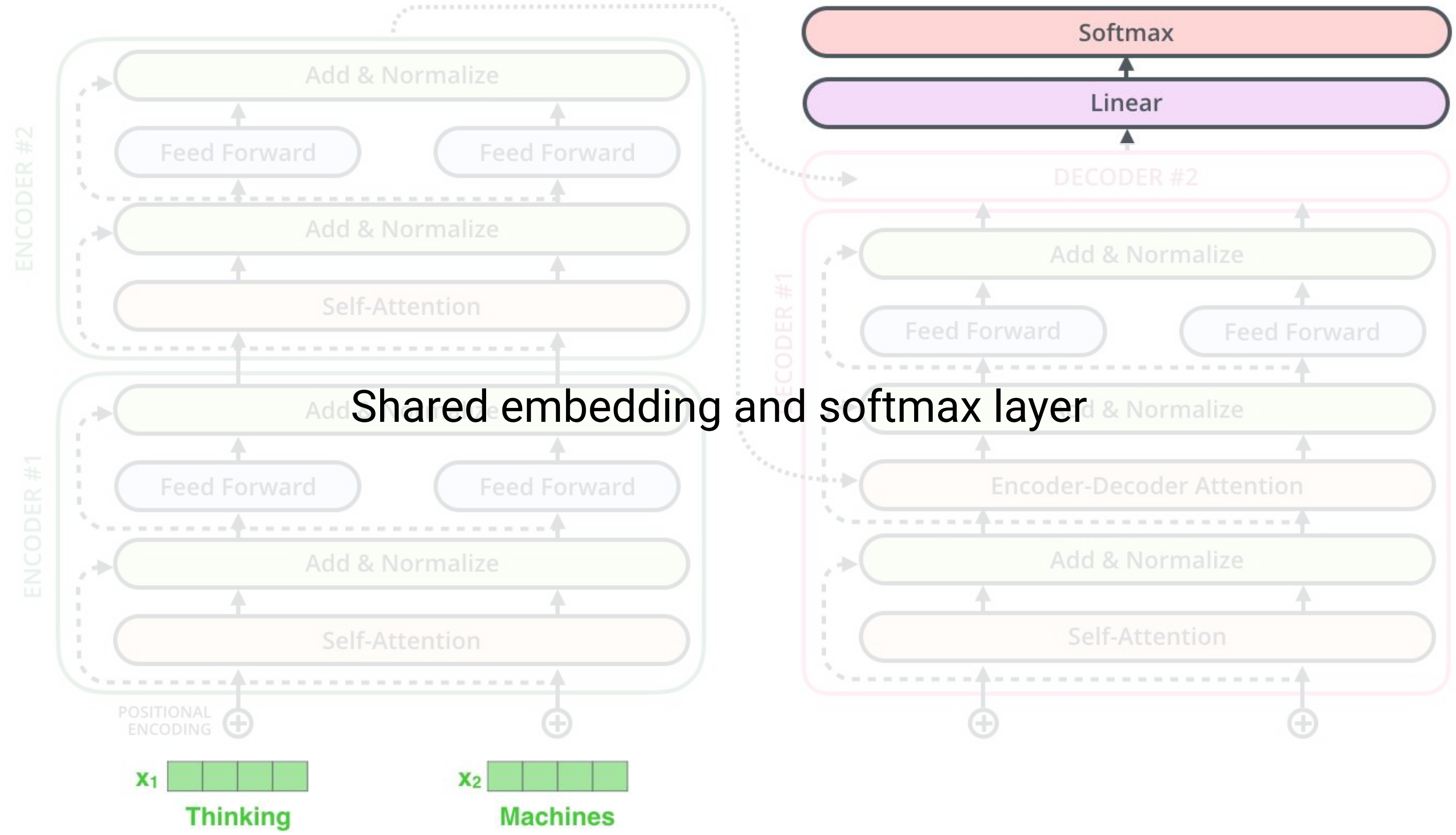


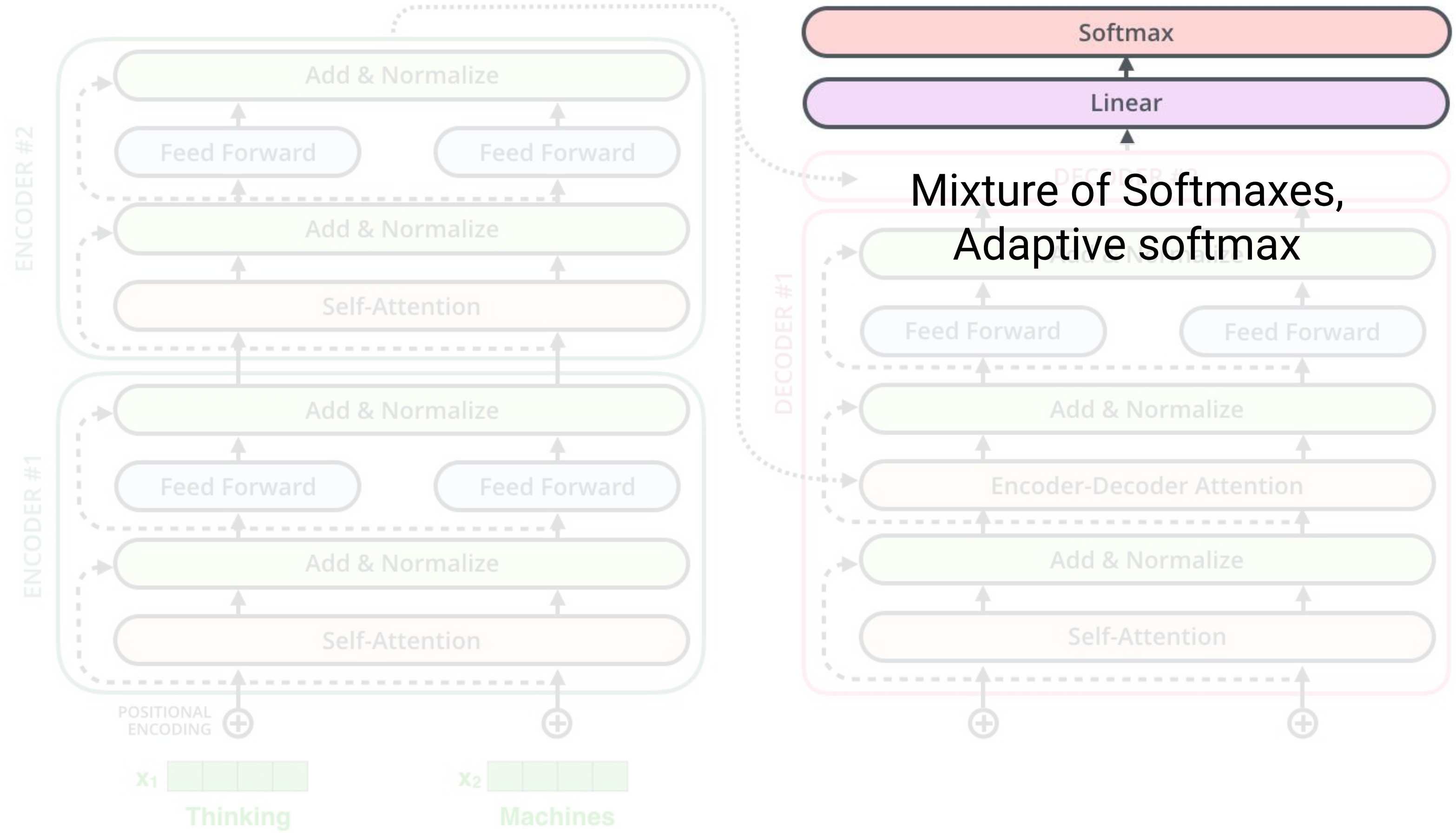
Source: http://jalammar.github.io/illustrated-transformer/



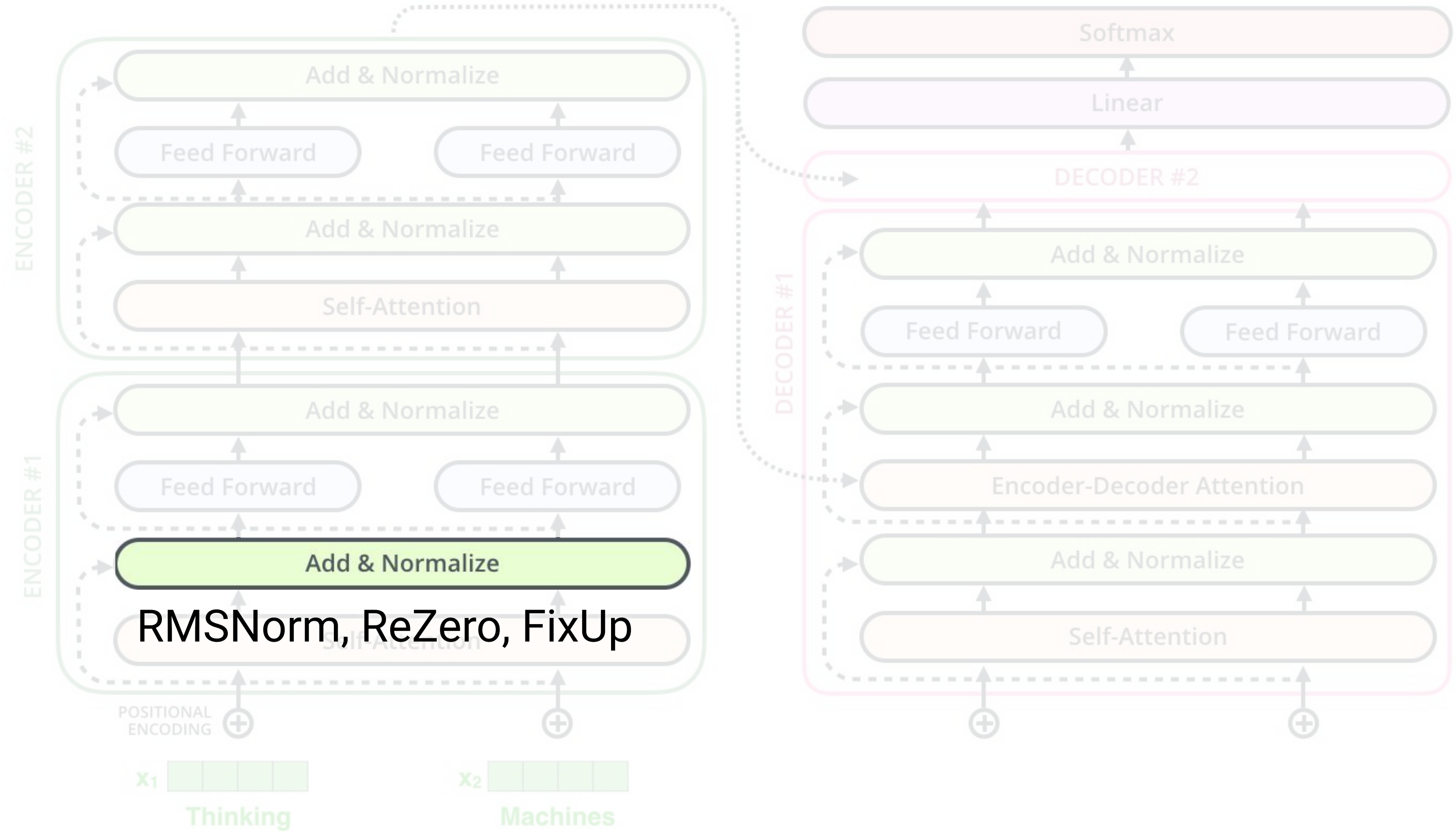
Factorized embeddings

Shared embedding and softmax layer

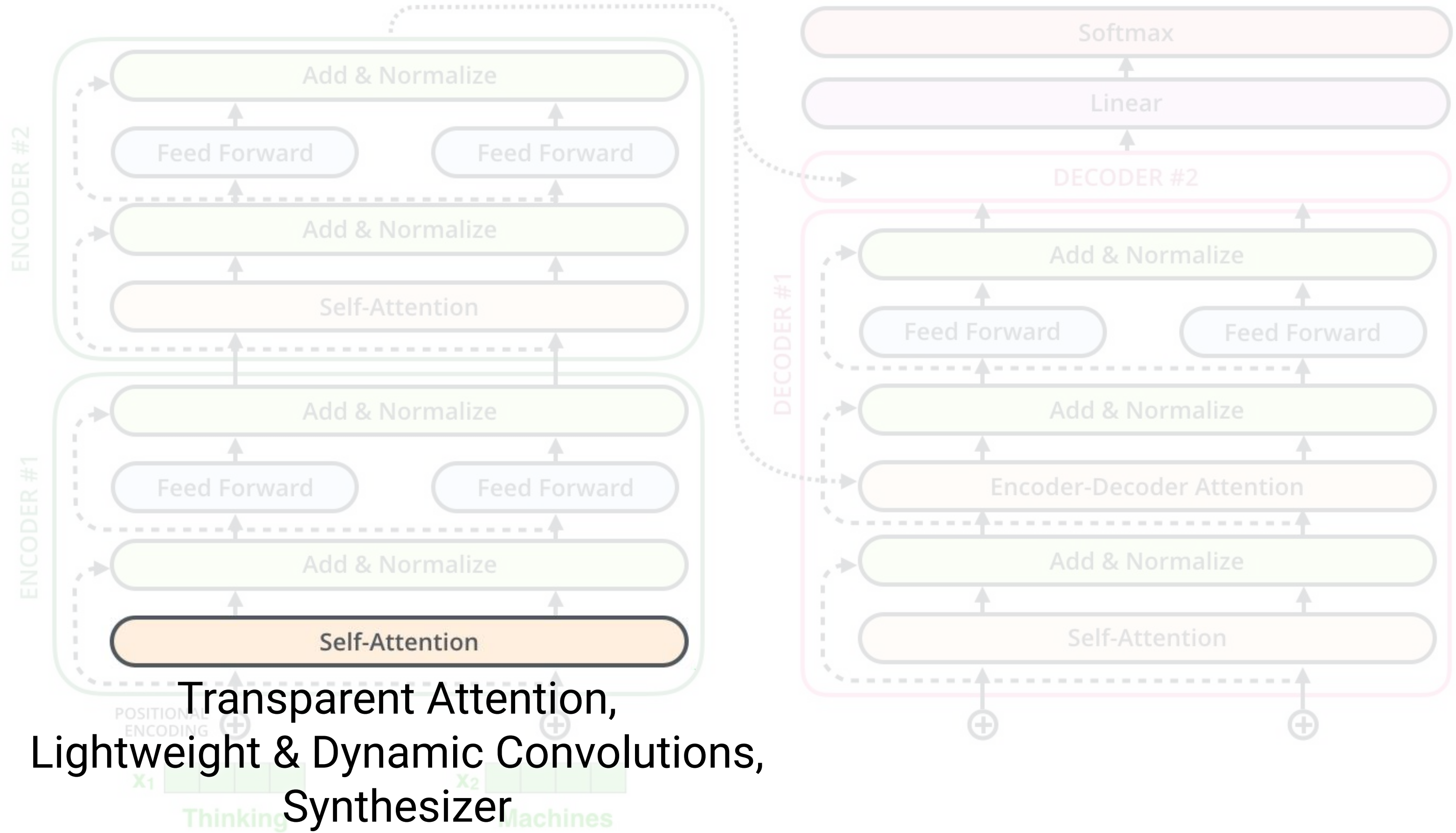




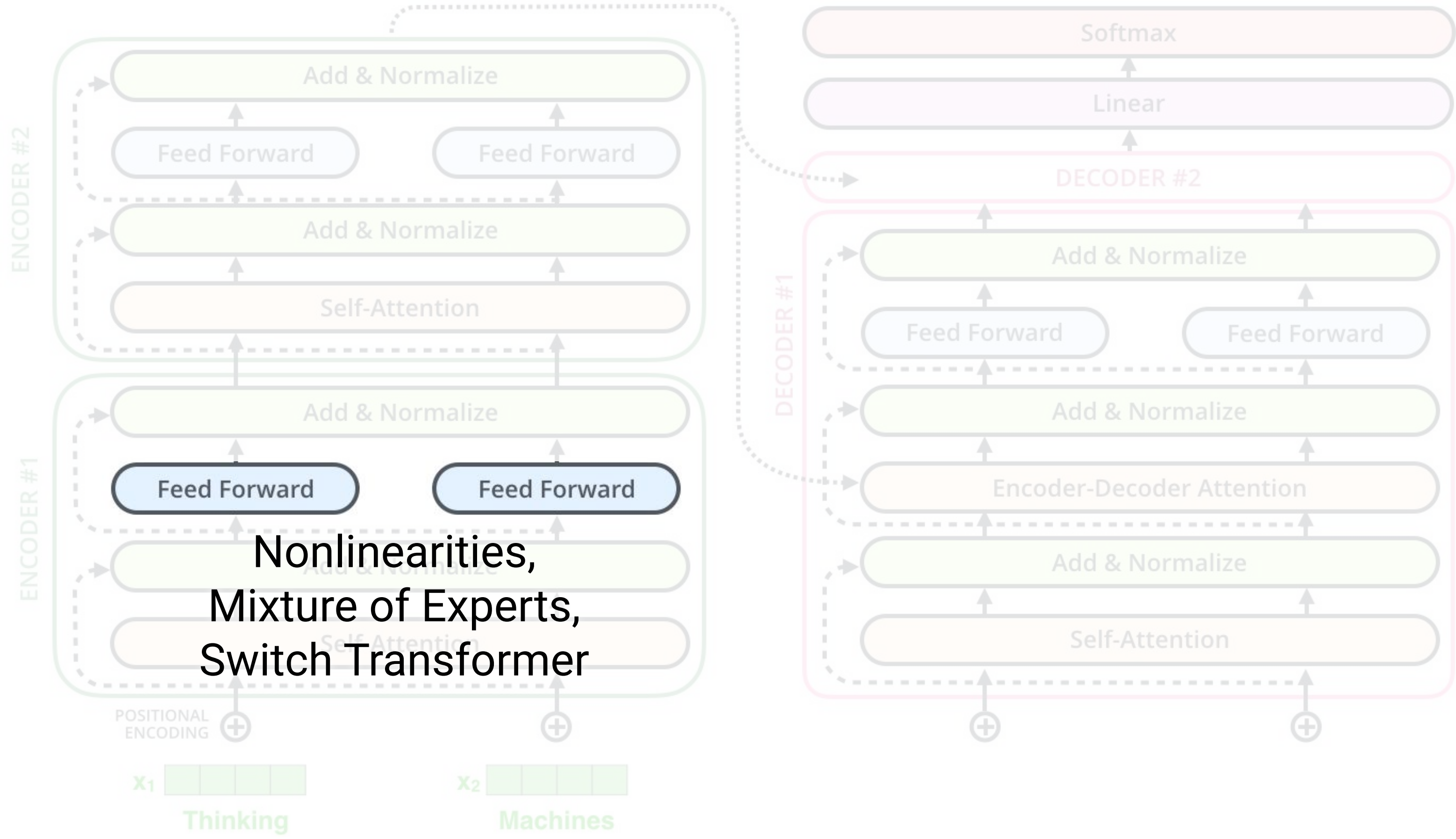
Mixture of Softmaxes,
Adaptive softmax



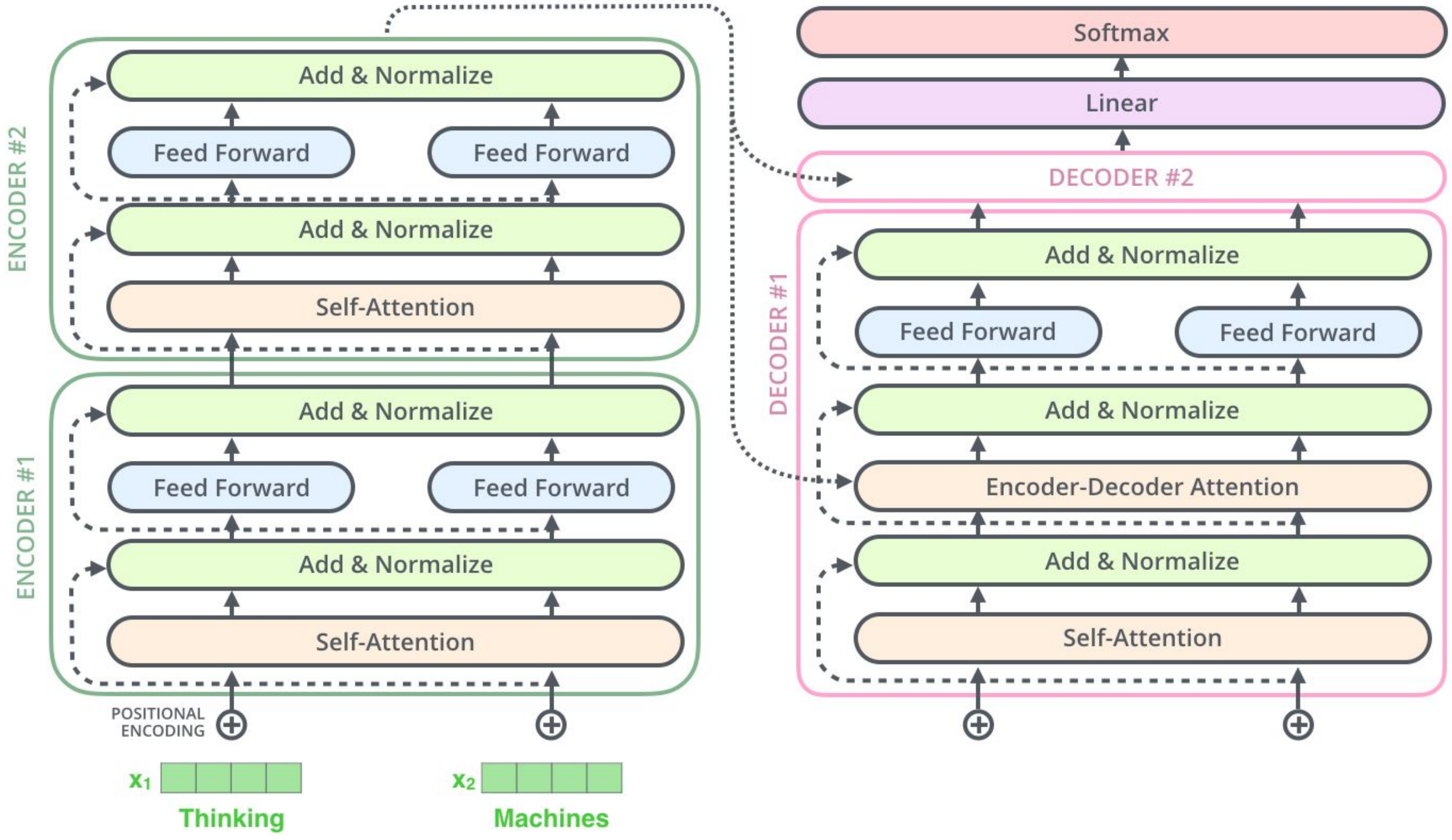
RMSNorm, ReZero, FixUp



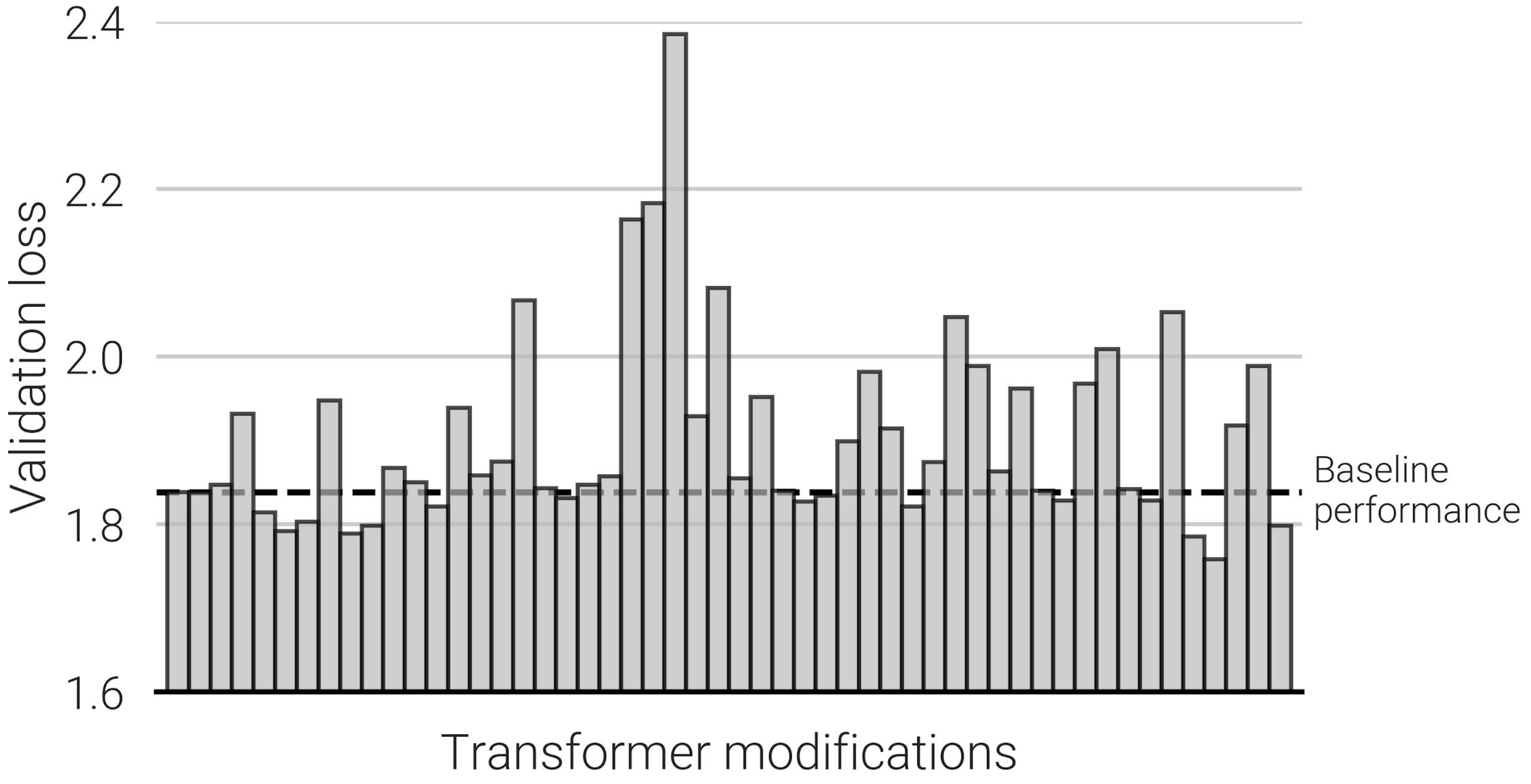
Transparent Attention,
Lightweight & Dynamic Convolutions,
Synthesizer

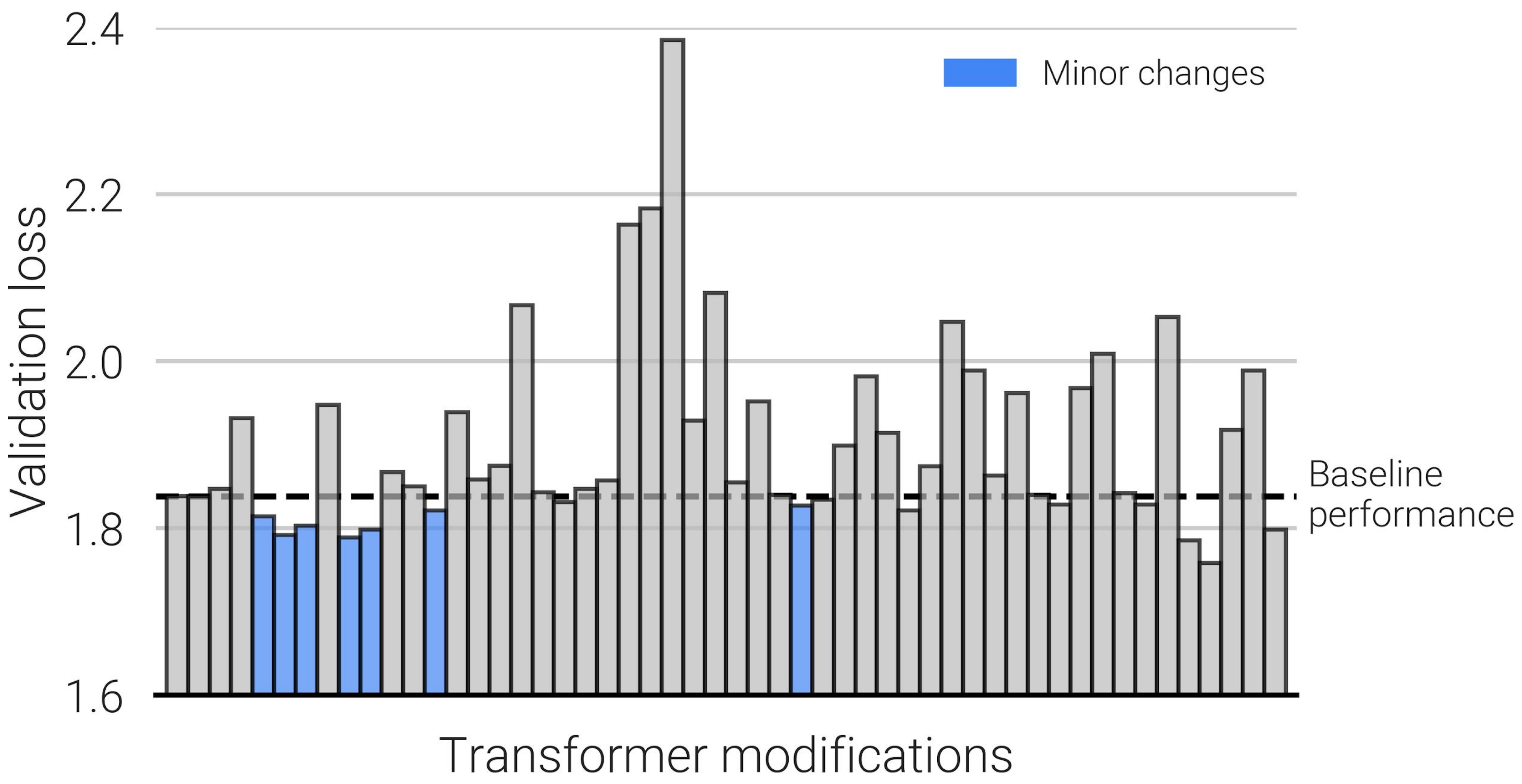


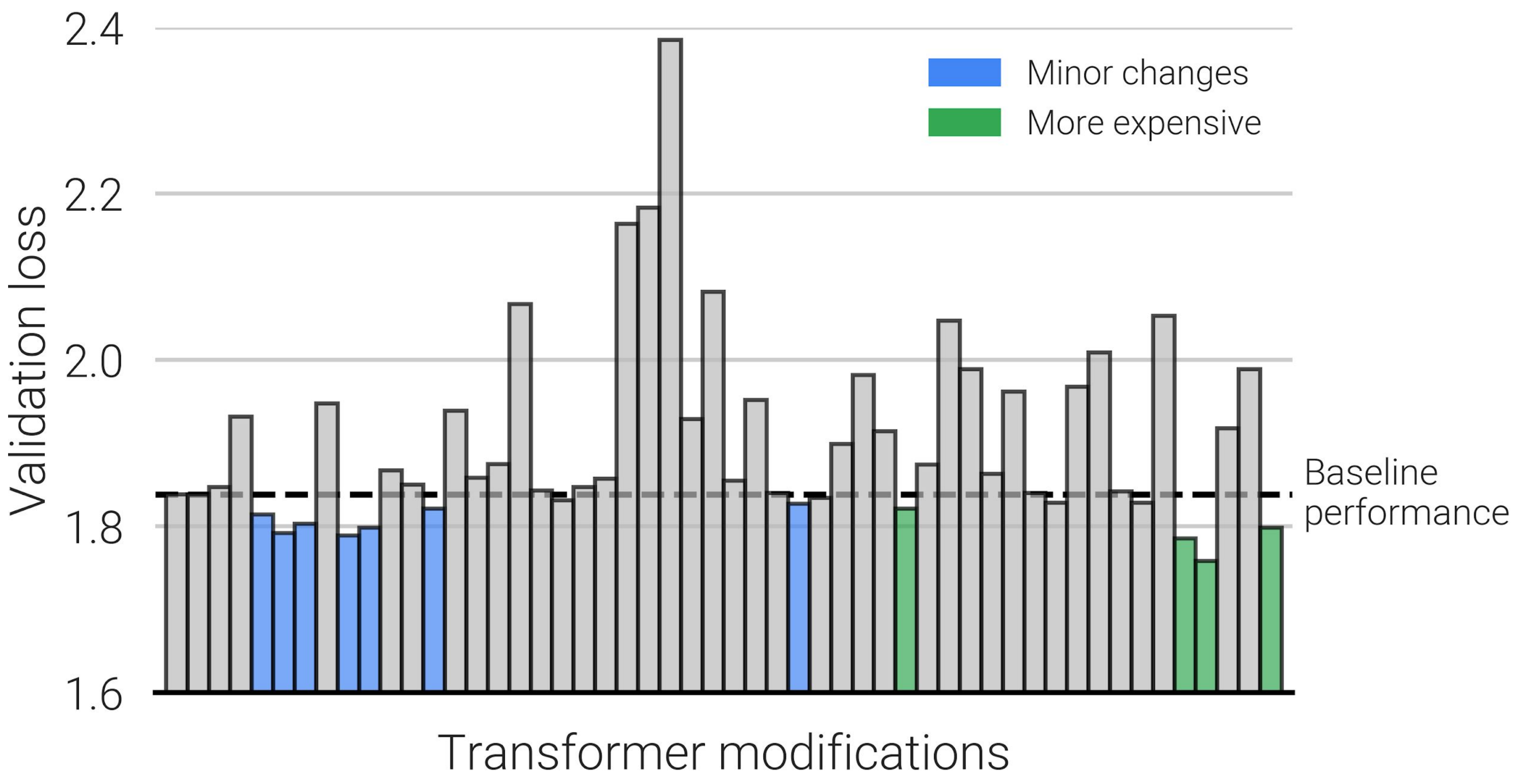
Nonlinearities,
Mixture of Experts,
Switch Transformer



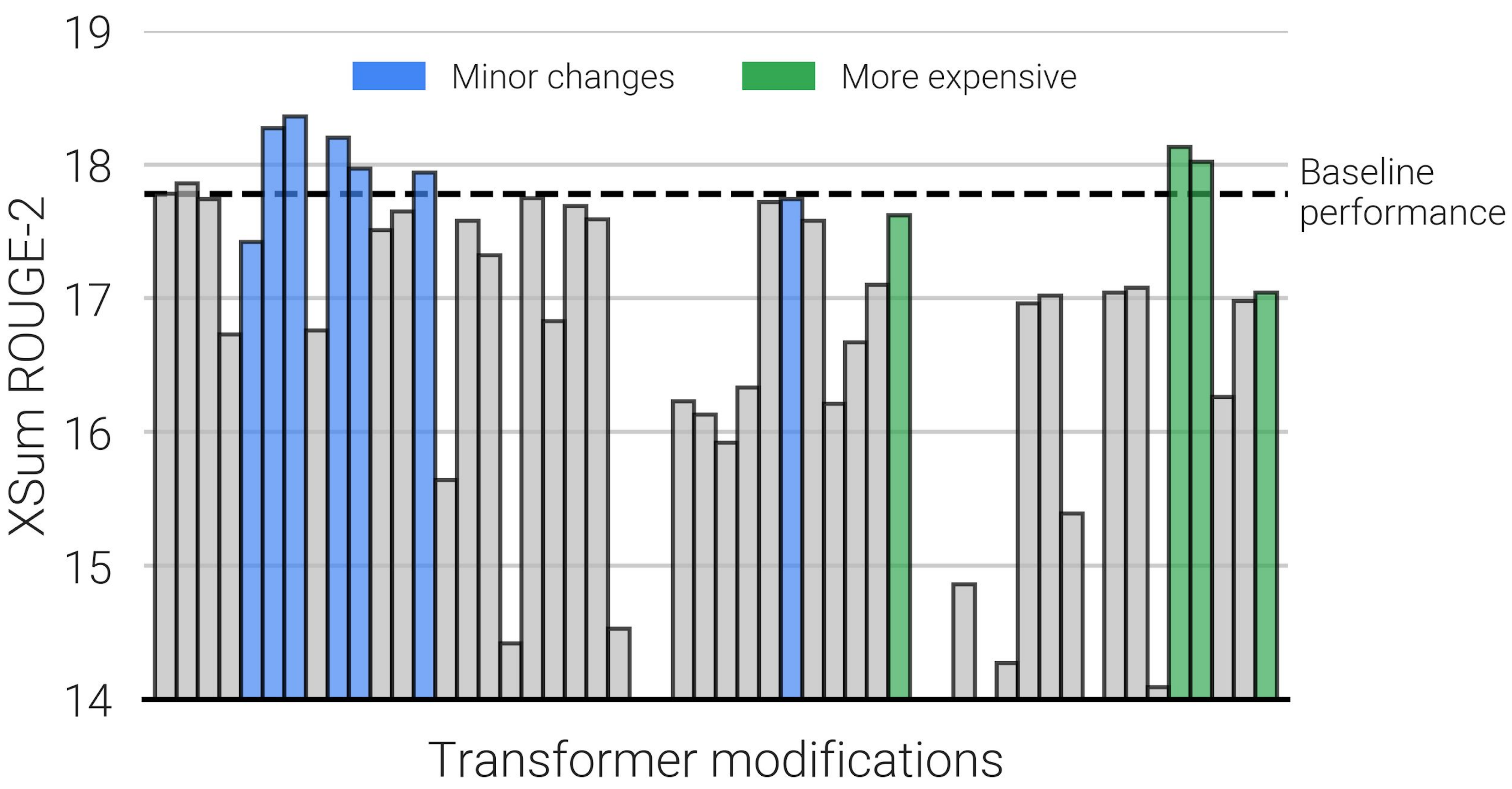
Funnel Transformer, Evolved Transformer, Universal Transformer, block sharing ...


2.4
2.2
2.0
1.8
1.6
Validation loss
Baseline performance
Transformer modifications

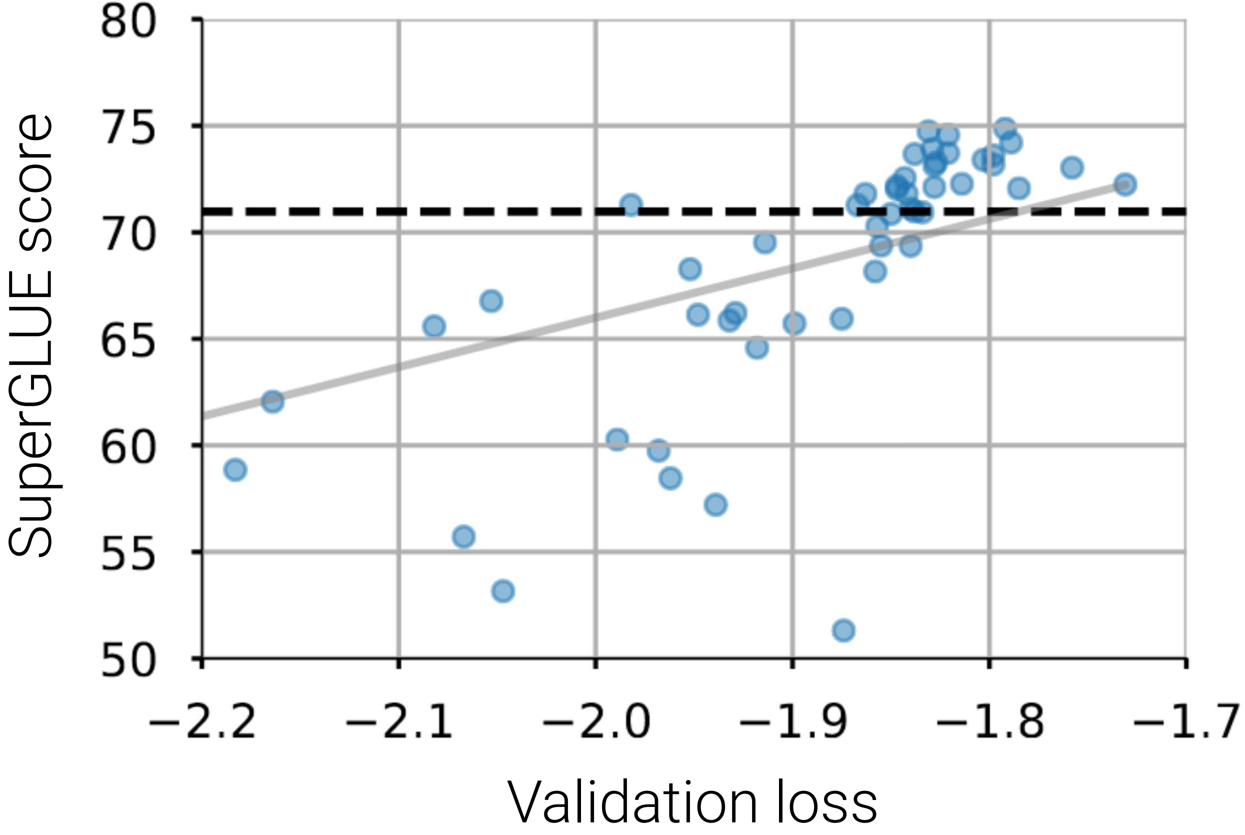

2.4
2.2
2.0
1.8
1.6
Validation loss
Minor changes
Baseline performance
Transformer modifications



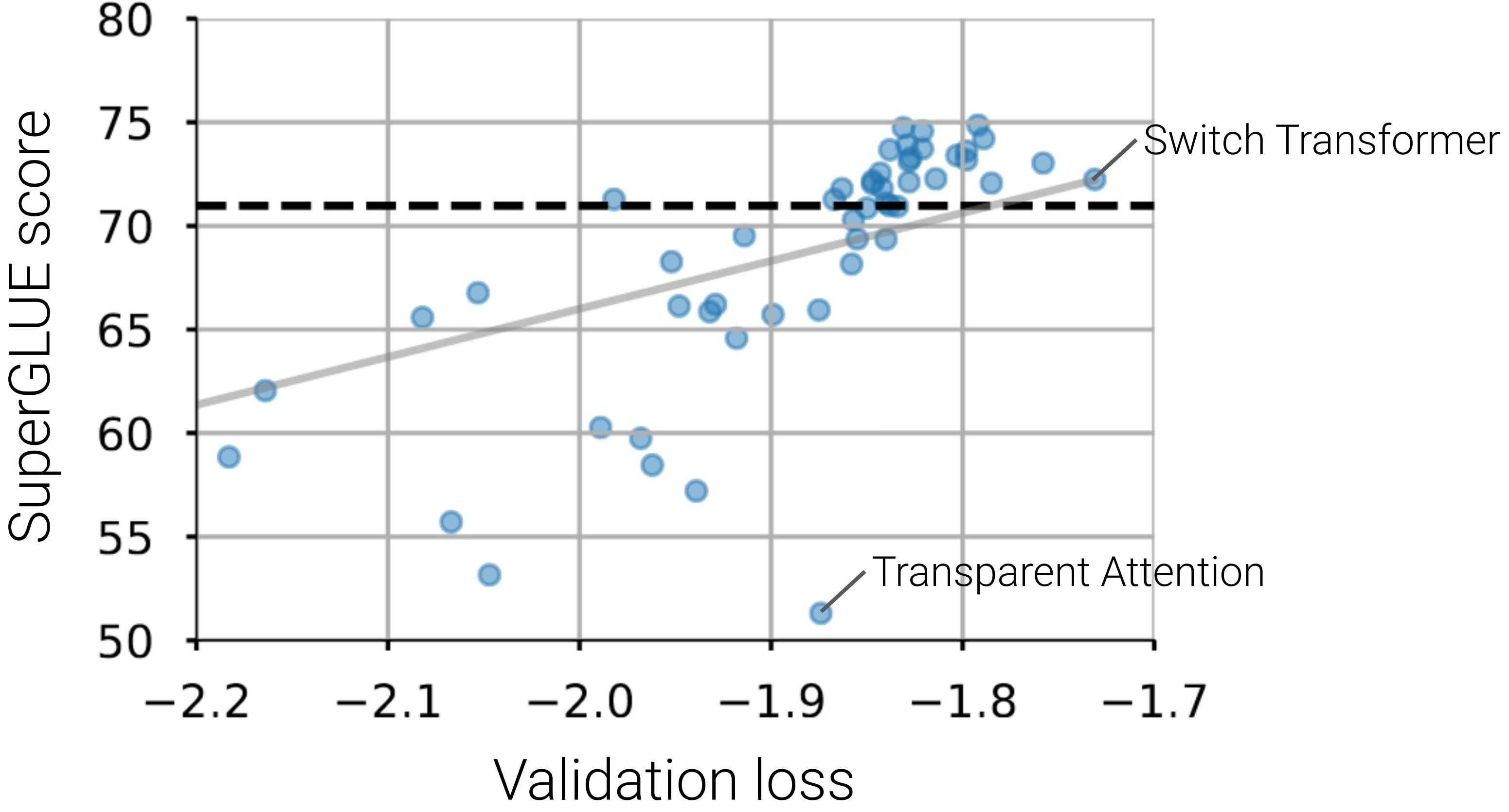
Validation loss
2.4
2.2
2.0
1.8
1.6
Minor changes
More expensive
Baseline performance
Transformer modifications



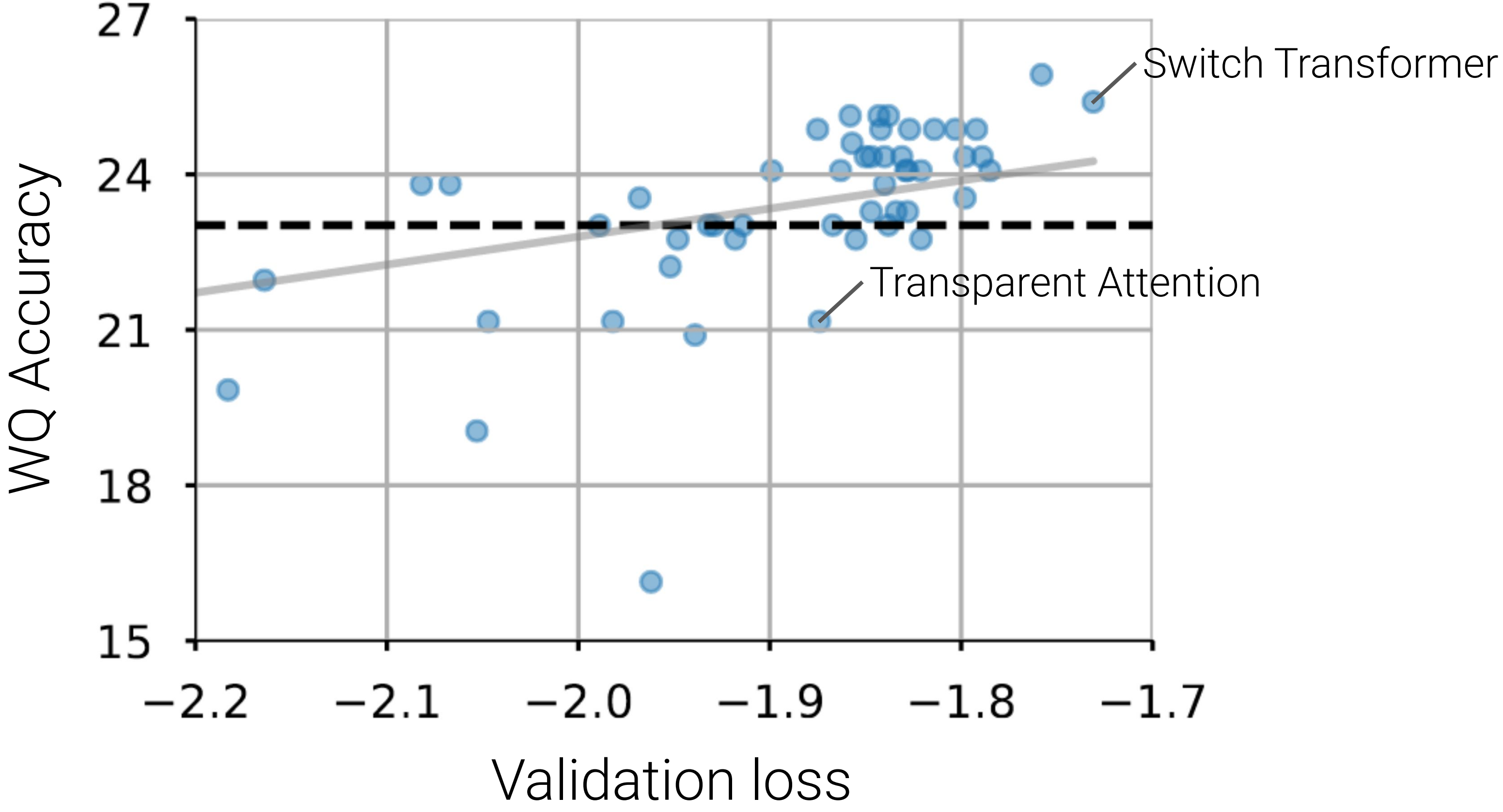
19
18
17
16
15
14
Minor changes
More expensive
Baseline performance
XSum ROUGE-2
Transformer modifications



80
75
70
65
60
55
50
SuperGLUE score
−2.2
−2.1
−2.0
−1.9
−1.8
−1.7
Validation loss



Switch Transformer
Transparent Attention
SuperGLUE score
Validation loss
80
75
70
65
60
55
50
−2.2
−2.1
−2.0
−1.9
−1.8
−1.7



Switch Transformer
Transparent Attention
WQ Accuracy
27
24
21
18
15
−2.2
−2.1
−2.0
−1.9
−1.8
−1.7
Validation loss

- Is our codebase unusual?
- Are our tasks non-standard?
- Do we need to tune hyperparameters?
- Did we implement the modifications correctly?
- Do Transformer modifications not “transfer”?

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

[mT5: A massively multilingual pre-trained text-to-text transformer]

[How Much Knowledge Can You Pack Into the Parameters of a Language Model?]

[Extracting Training Data from Large Language Models]

[Do Transformer Modifications Transfer Across Implementations and Applications?]

Work done with Adam Roberts, Aditya Barua, Aditya Siddhant, Alina Oprea, Ariel Herbert-Voss, Dawn Song, Eric Wallace, Florian Tramer, Hyung Won Chung, Jake Marcus, Karishma Malkan, Katherine Lee, Linting Xue, Matthew Jagielski, Michael Matena, Mihir Kale, Nan Ding, Nicholas Carlini, Noah Constant, Noah Fiedel, Noam Shazeer, Peter J. Liu, Rami Al-Rfou, Sharan Narang, Thibault Fevry, Tom Brown, Ulfar Erlingsson, Wei Li, William Fedus, Yanqi Zhou, Yi Tay, and Zhenzhong Lan

Questions?