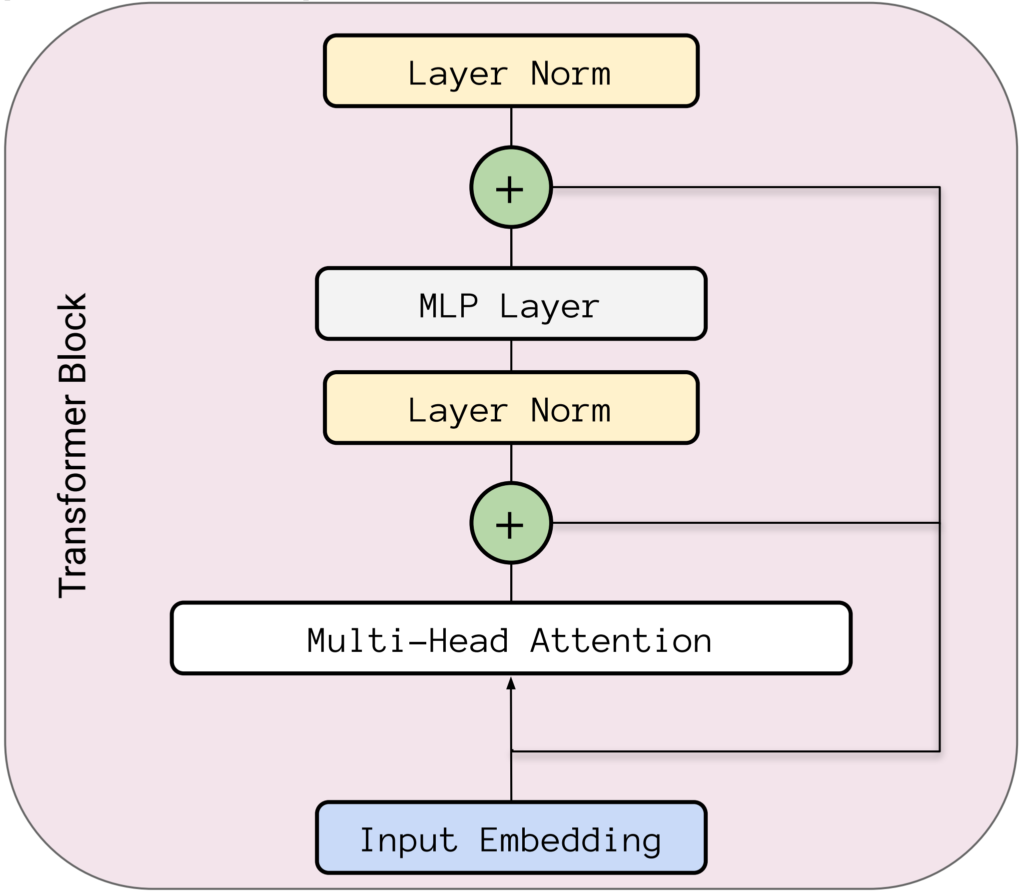
Training Large Language Models

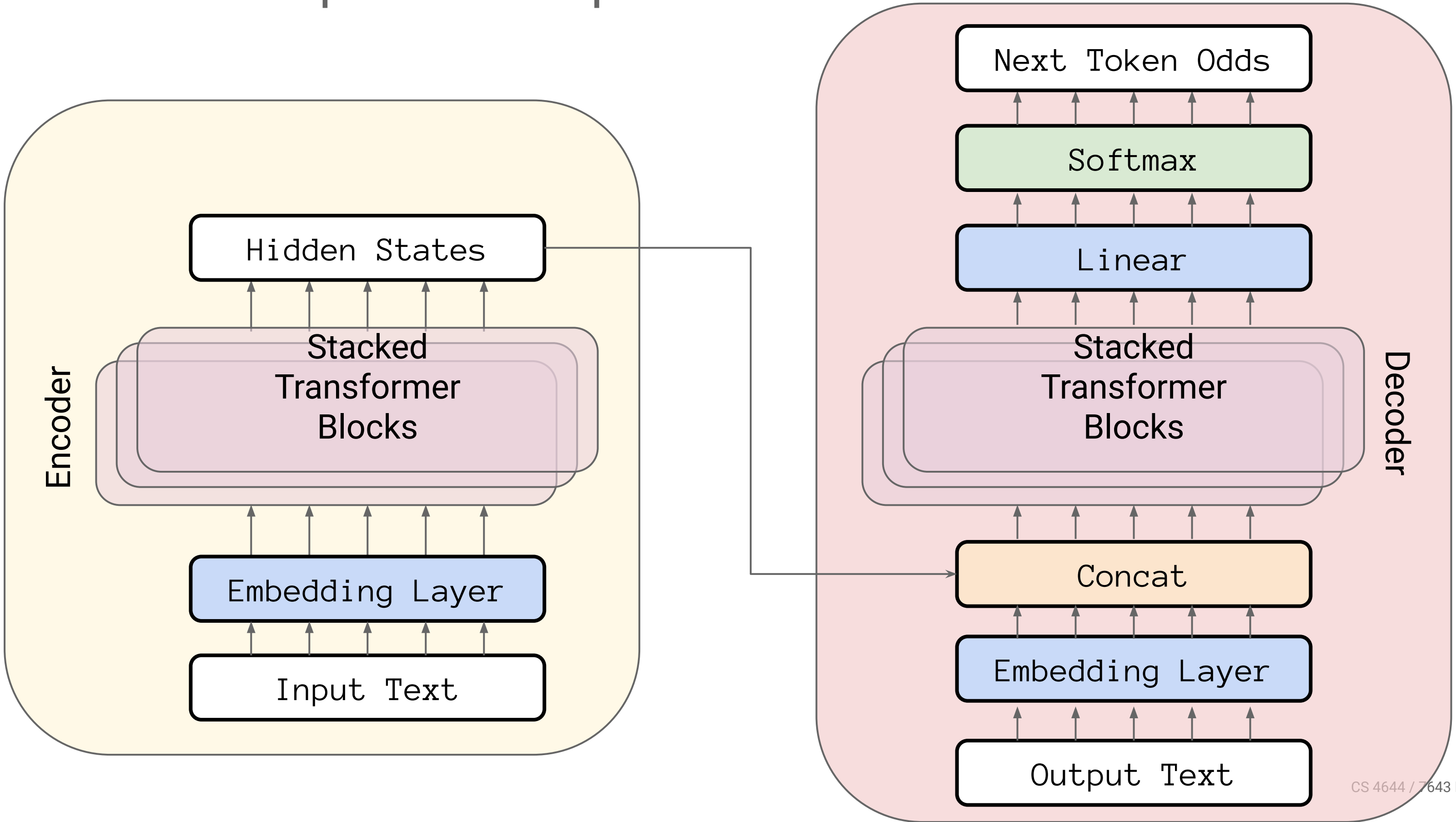
CS 4644 / 7643: Deep Learning

William Held
School of Interactive Computing
Georgia Institute of Technology

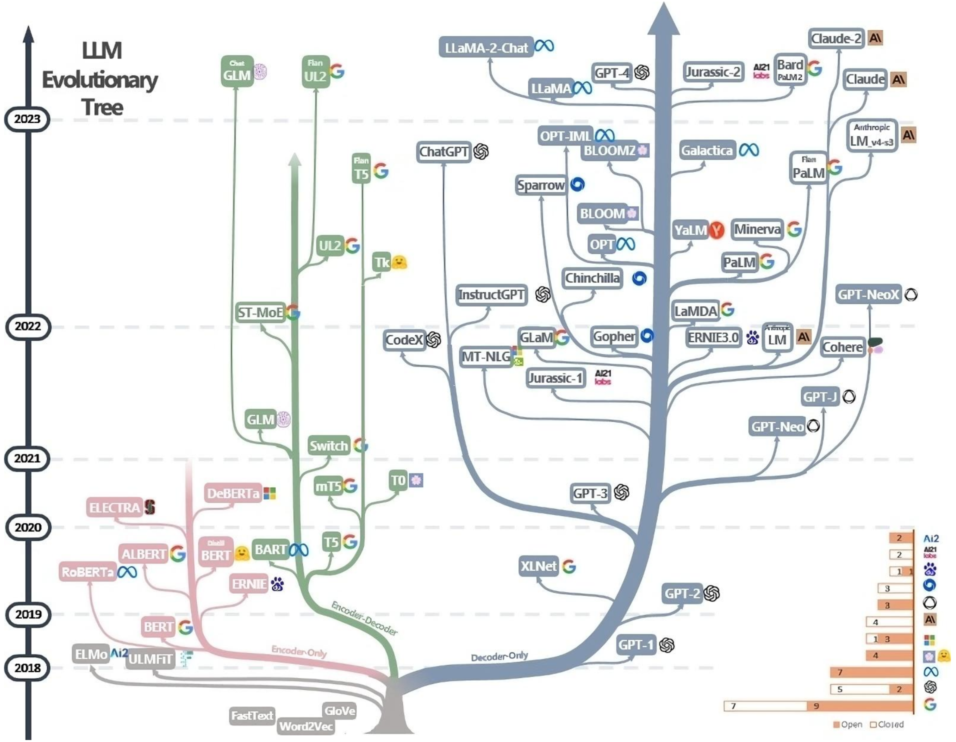
Last Lecture Speed Recap: The Transformer Block



Last Lecture Speed Recap: Attention is “All” You Need



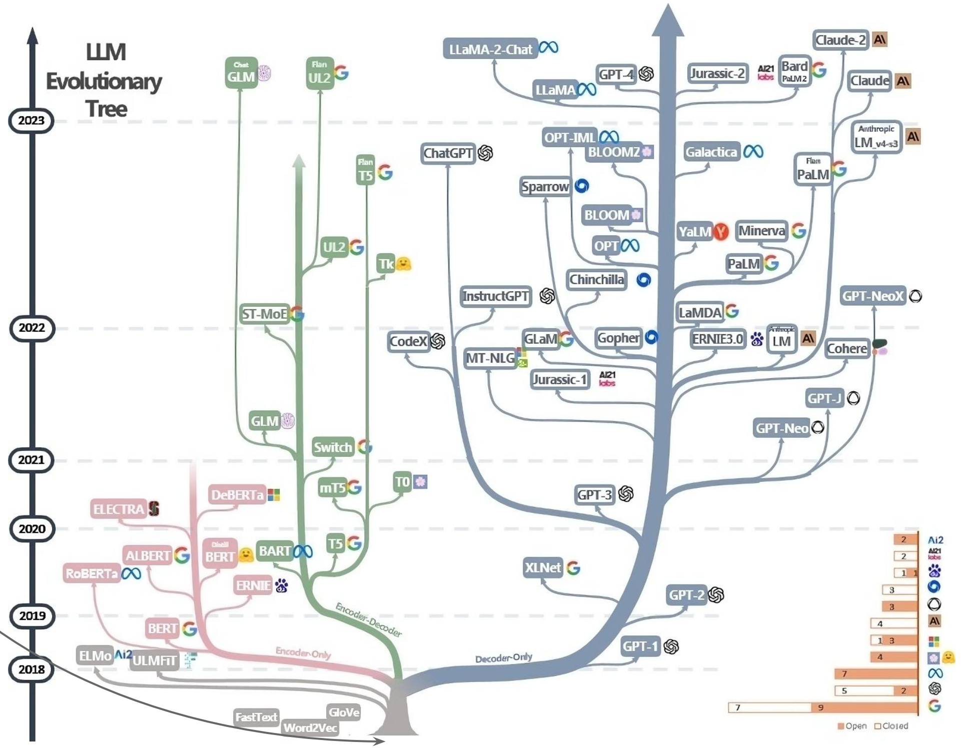
How do we go from purpose driven models to LLMs?



<https://github.com/Mooler0410/LLMsPracticalGuide>

How do we go from purpose driven models to LLMs?

Self-Supervised Learning
How do we most effectively turn raw text into meaningful loss?



https://github.com/Mooler0410/LLMsPracticalGuide

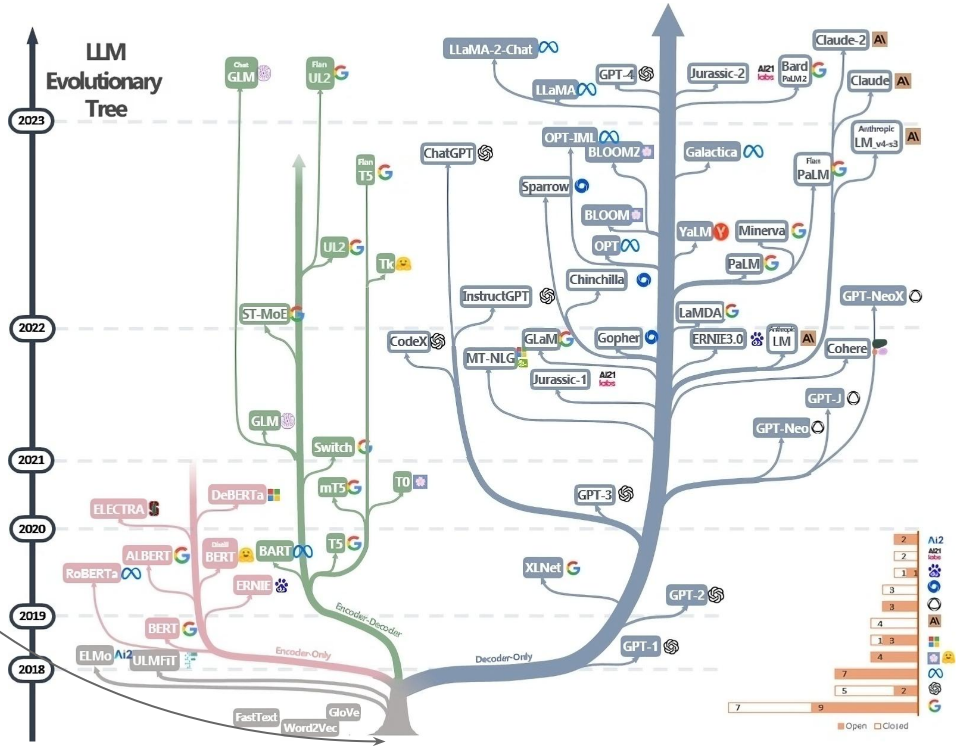
How do we go from purpose driven models to LLMs?

Self-Supervised Learning

How do we most effectively turn raw text into meaningful loss?

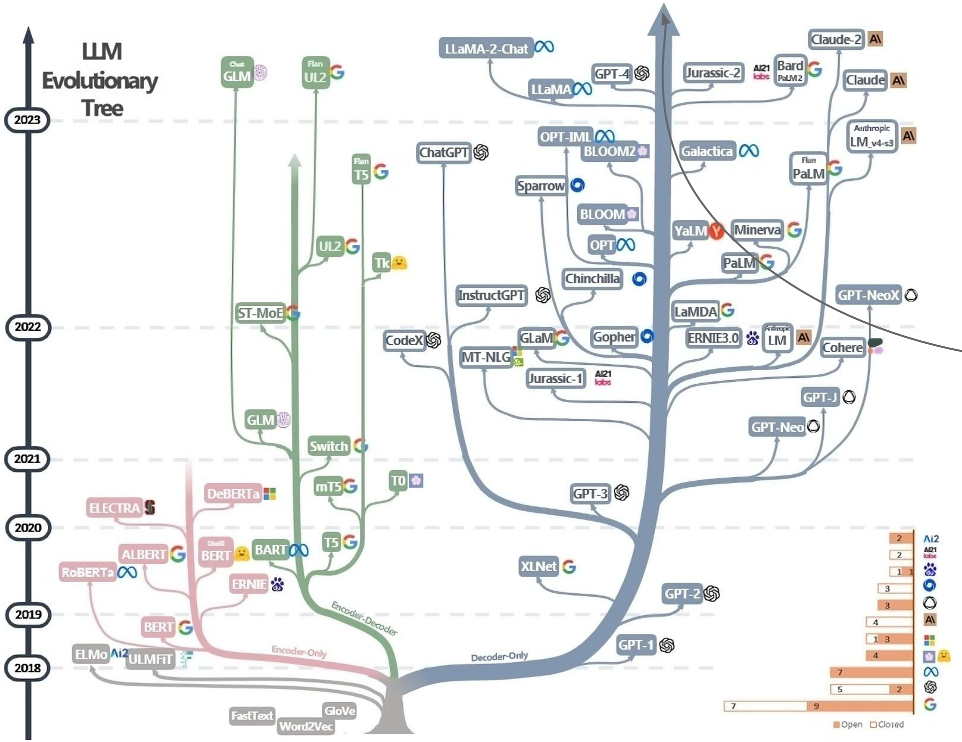
<u>Covered Today</u>

- Encoder Only
- Decoder Only
- Encoder-Decoder



https://github.com/Mooler0410/LLMsPracticalGuide

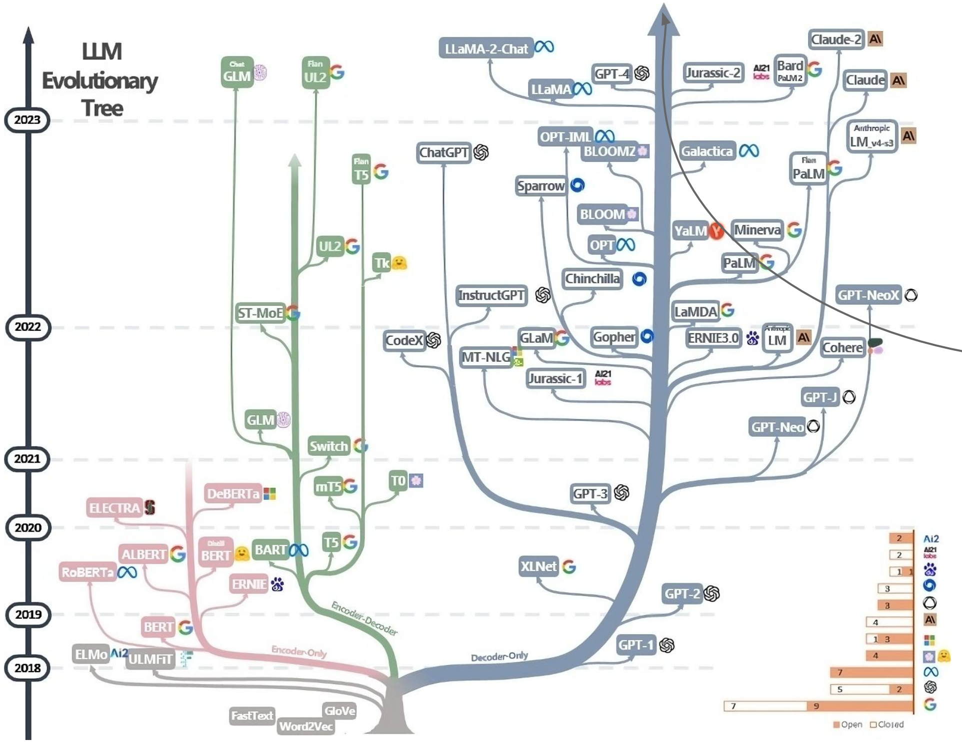
How do we go from purpose driven models to LLMs?



Data Scaling
How do we source and train on high-quality data at scale?

https://github.com/Mooler0410/LLMsPracticalGuide

How do we go from purpose driven models to LLMs?



Data Scaling

How do we source and train on high-quality data at scale?

Covered Today

- Data Curation Over Time
- Distributed Training

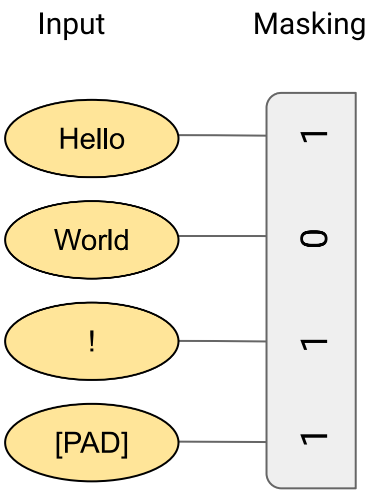
https://github.com/Mooler0410/LLMsPracticalGuide

LLM Advancements have been driven primarily by these two

Self-Supervised Learning
How do we most effectively turn raw text into meaningful loss?

Data Scaling
How do we source and train on high-quality data at scale?

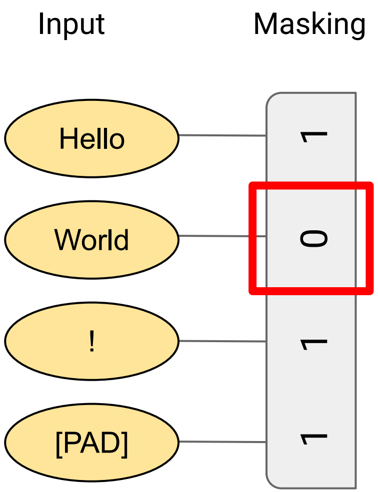
SSL | From raw text to loss!



Masked Language Model

[Devlin et al. 2018 (BERT)]

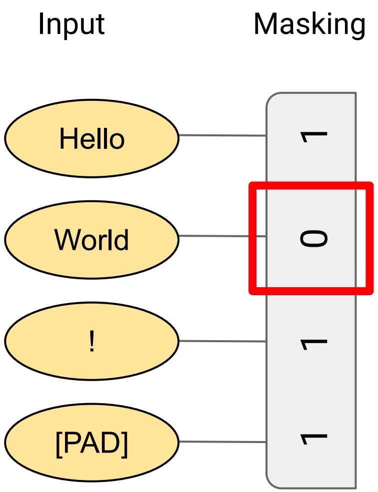
SSL | What is the "Mask" in a Masked Language Model?



Masked Language Model

[Devlin et al. 2018 (BERT)]

SSL | What is the “Mask” in a Masked Language Model?



Recall

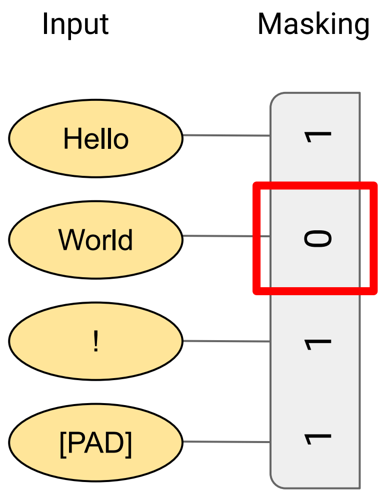
```
Similarities: E = QXT / sqrt(DQ)
Attention Matrix: A = softmax(E,dim=1)
Output vectors: Y = AX
```

$Y_i = \sum_j A_{i,j} X$

Masked Language Model

[Devlin et al. 2018 (BERT)](#)

SSL | What is the “Mask” in a Masked Language Model?



Masked Attention

Similarities: $E = (QX^T / \text{sqrt}(D_Q)) * \text{MASK}$

Attention Matrix: $A = \text{softmax}(E, \text{dim}=1)$

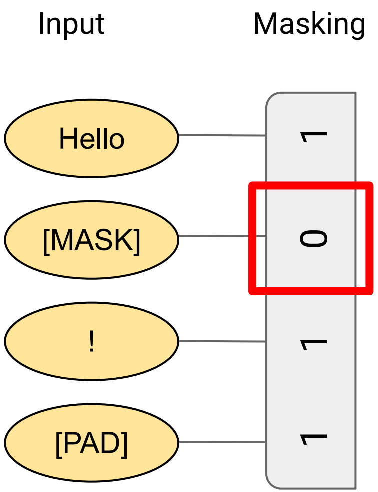
Output vectors: $Y = AX$

$Y_i = \sum_j A_{i,j} X$

Masked Language Model

[Devlin et al. 2018 (BERT)]

SSL | What is the “Mask” in a Masked Language Model?



Intuition

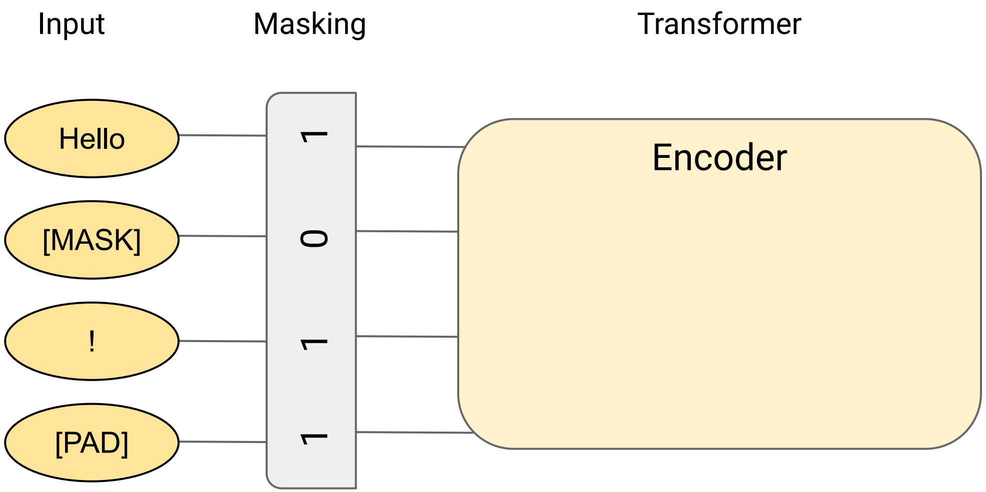
If $MASK_i = 0$, then $Y_i = \sum_{j, j != i} A_{i,j} X$

a.k.a the representation of the masked token is created purely from context

Masked Language Model

[Devlin et al. 2018 (BERT)](#)

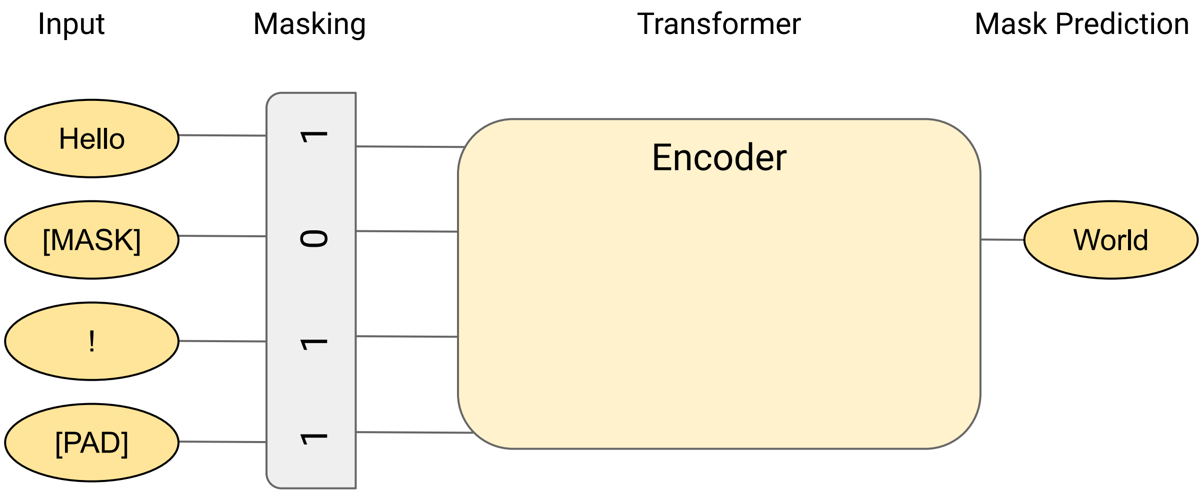
SSL | Masked Token Prediction



Masked Language Model

[Devlin et al. 2018 (BERT)]

SSL | Masked Token Prediction



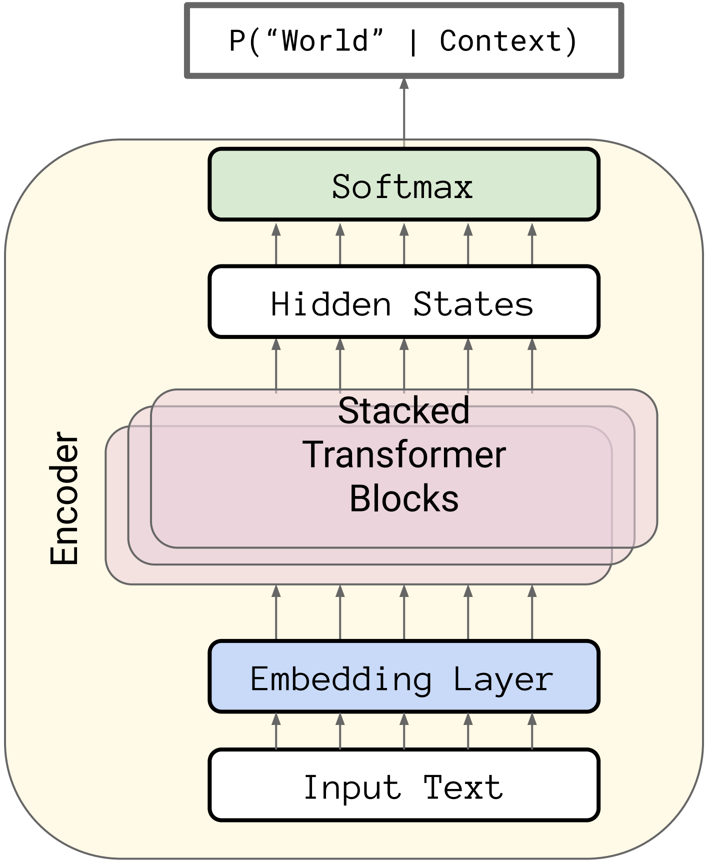
Masked Language Model

[Devlin et al. 2018 (BERT)]

SSL | Masked Token Prediction

Optimize Negative Log Likelihood

`loss = -log(P("World" | Context))`

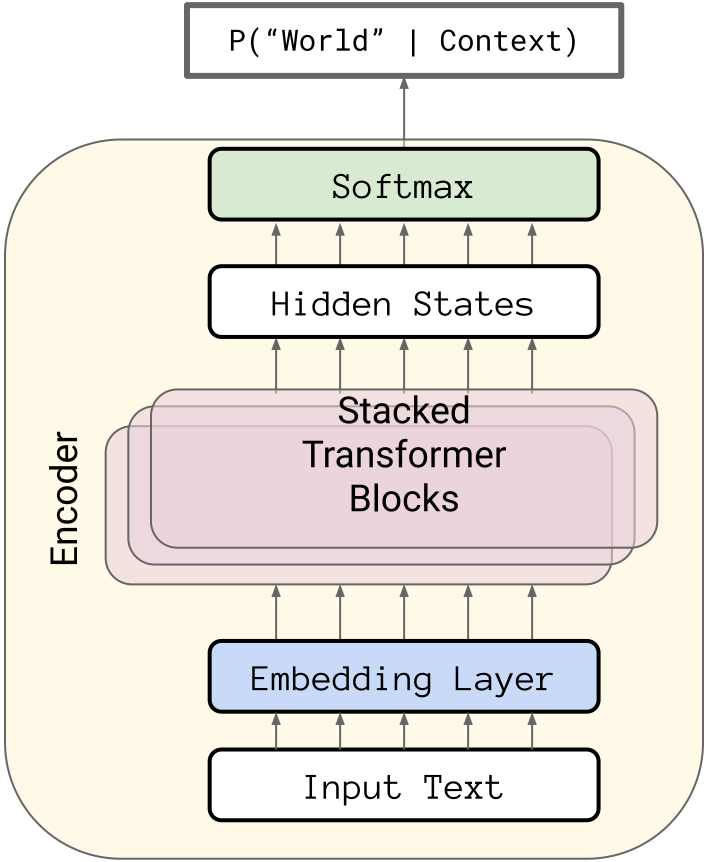


SSL | Masked Token Prediction

Optimize Negative Log Likelihood

`loss = -log(P("World" | Context))`

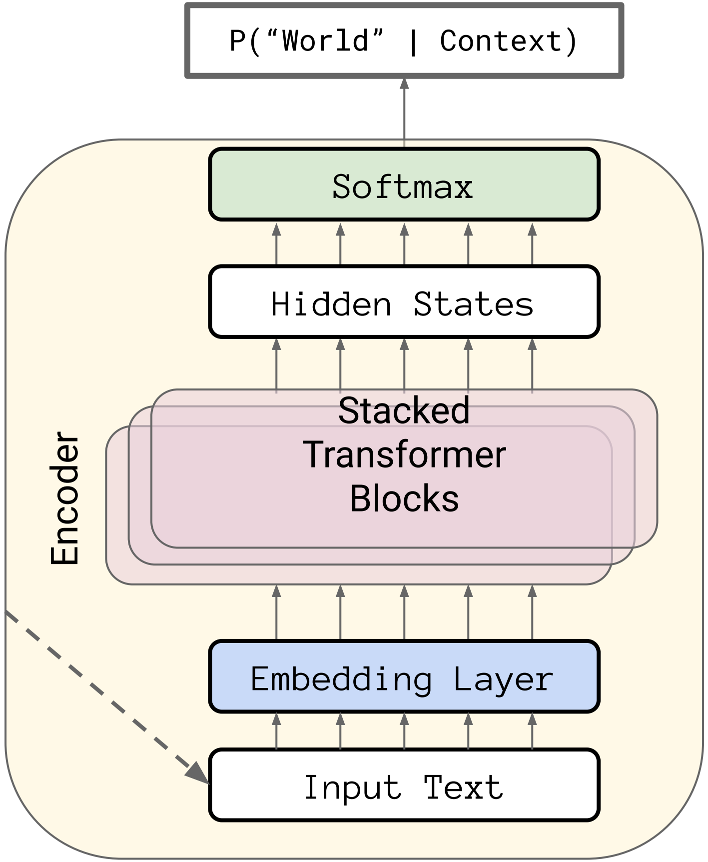
Equivalent to the Cross-Entropy Loss term from Lecture 3!



Side Note | Tokens v.s. Words

Languages have a lot of words!

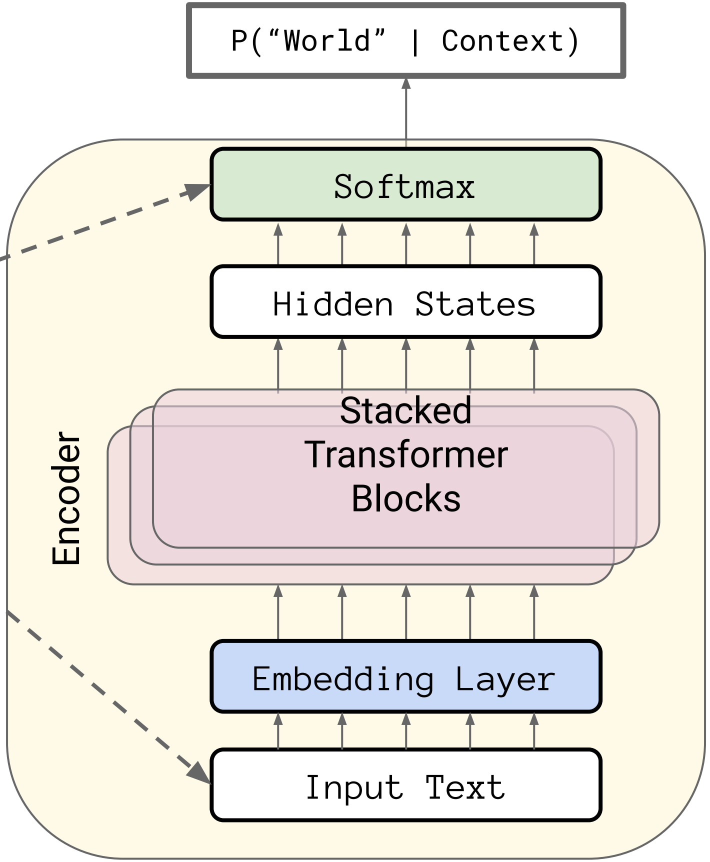
```
If V = Number of Words:
        O(V) Memory Scaling
```



Side Note | Tokens v.s. Words

Languages have a lot of words!

```
If V = Number of Words:
            O(V) Memory Scaling
           O(V) Runtime Scaling
```



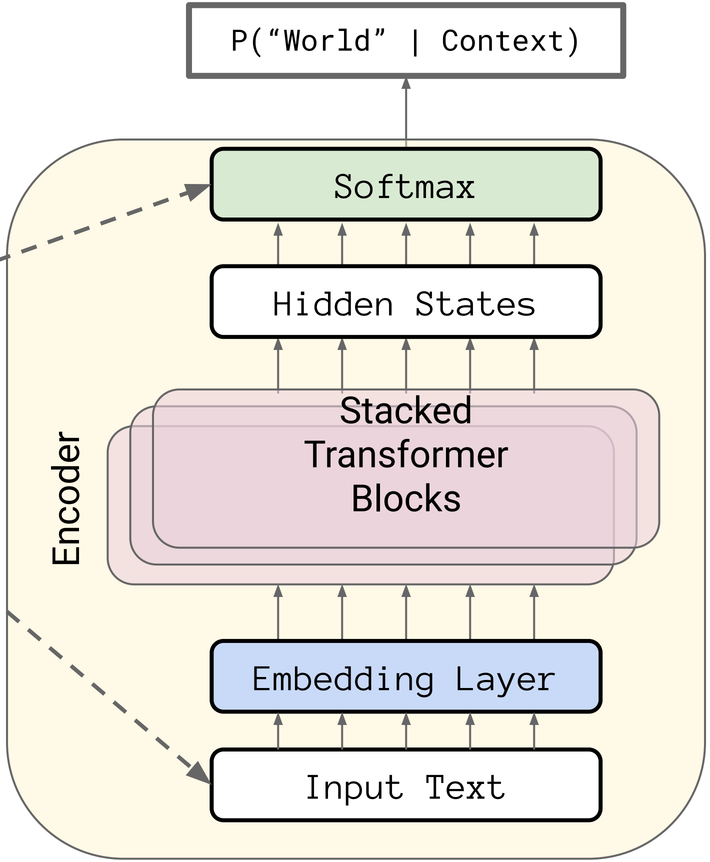
Side Note | Tokens v.s. Words

Languages have a lot of words!

```
If V = Number of Words:
        O(V) Memory Scaling
        O(V) Runtime Scaling
```

This limits our vocabulary size a lot.

```
Tokenizers:
   Pre-processing to split words into
smaller chunks called "Tokens" so that we
   can cover all words with smaller V
```



Side Note | Tokens v.s. Words

Languages have a lot of words!

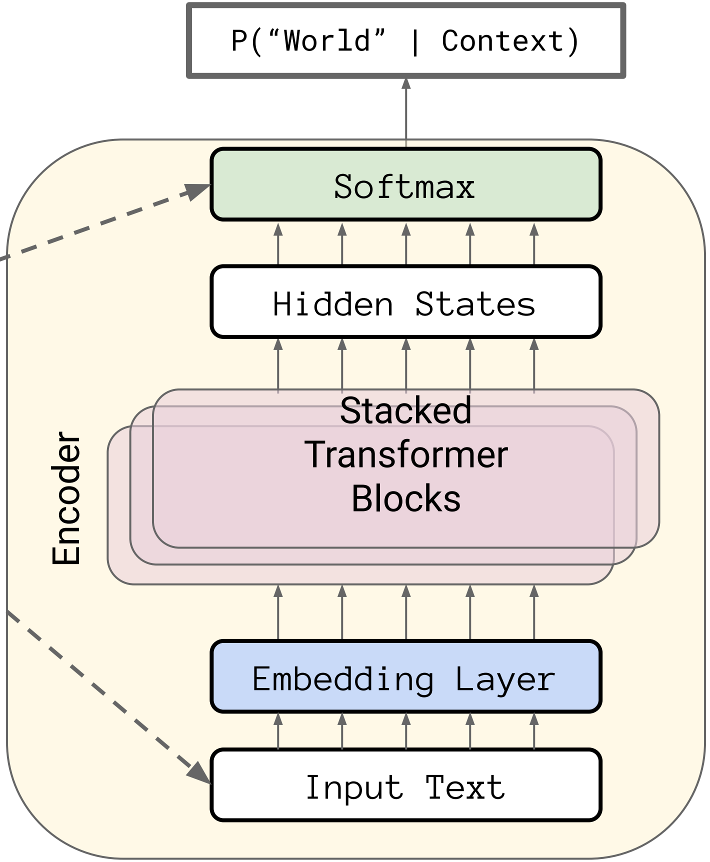
```
If V = Number of Words:
          O(V) Memory Scaling
          O(V) Runtime Scaling
```

This limits our vocabulary size a lot.

```
Tokenizers:
   Pre-processing to split words into
smaller chunks called "Tokens" so that we
   can cover all words with smaller V

  Important but outside of Course Scope
```

[HuggingFace Tokenizer Summary]



Data | BERT used existing curation!

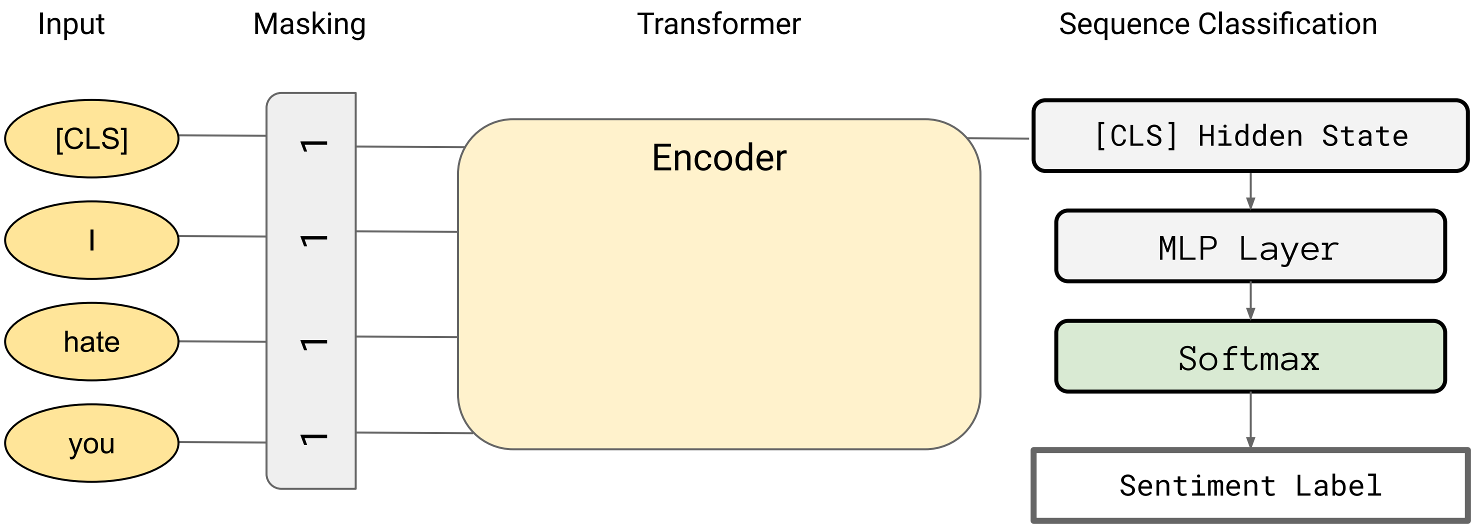
<u>BERT Corpus</u>
English Wikipedia + BooksCorpus

<u>Size</u>
~3 Billion Tokens

<u>Quality</u>
High quality text,
Broad “Academic” Knowledge,
Limited Diversity

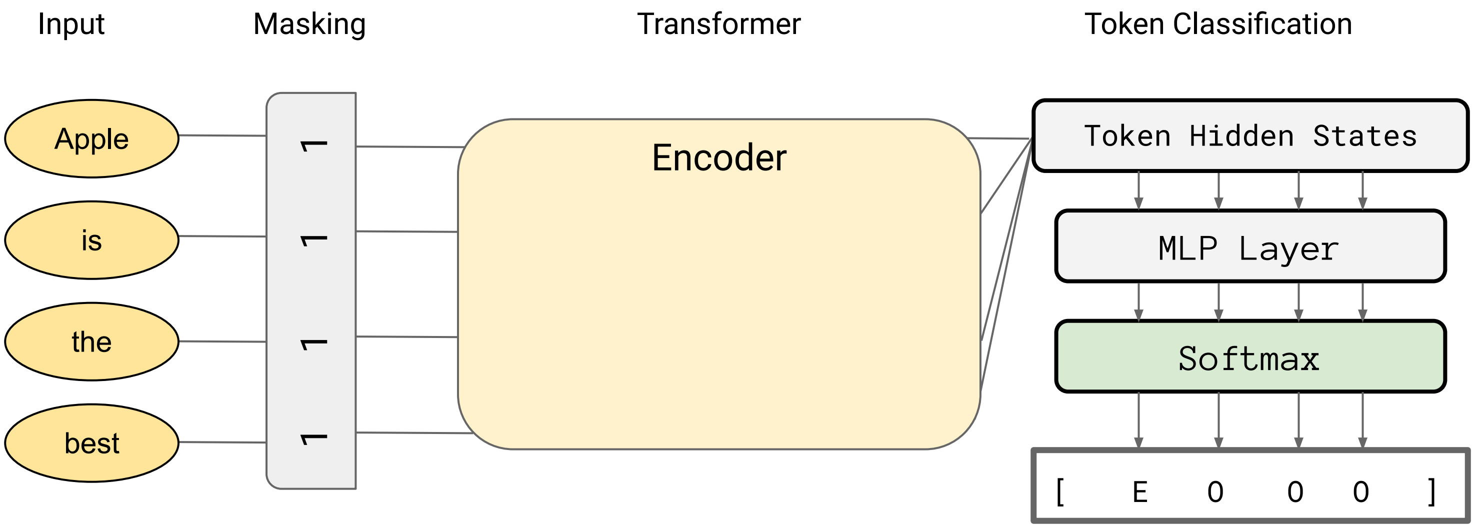
[Devlin et al. 2018 (BERT)]

Applications | Encoders as "Foundation" Language Models



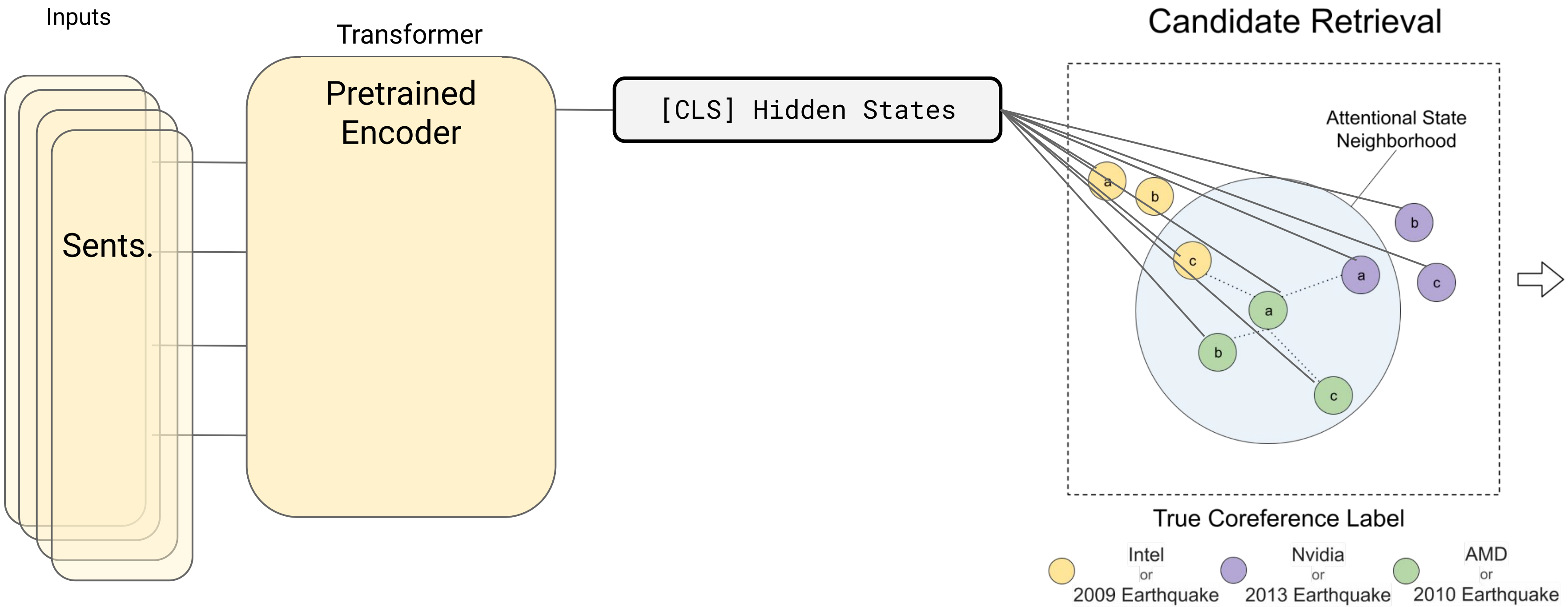
[Devlin et al. 2018 (BERT)](#)

Applications | Encoders as “Foundation” Language Models



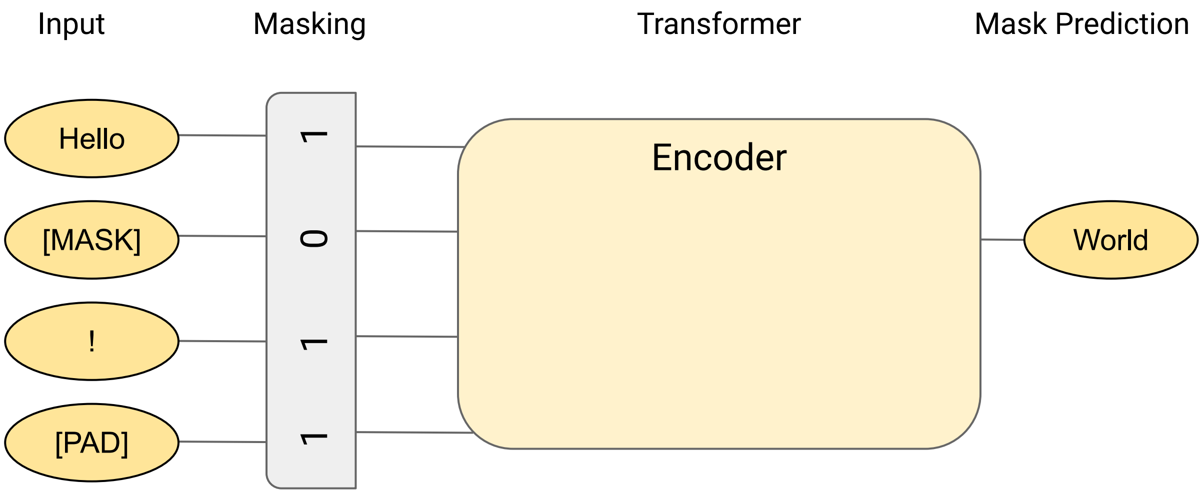
[Devlin et al. 2018 (BERT)]

Applications | Encoders as “Foundation” Language Models



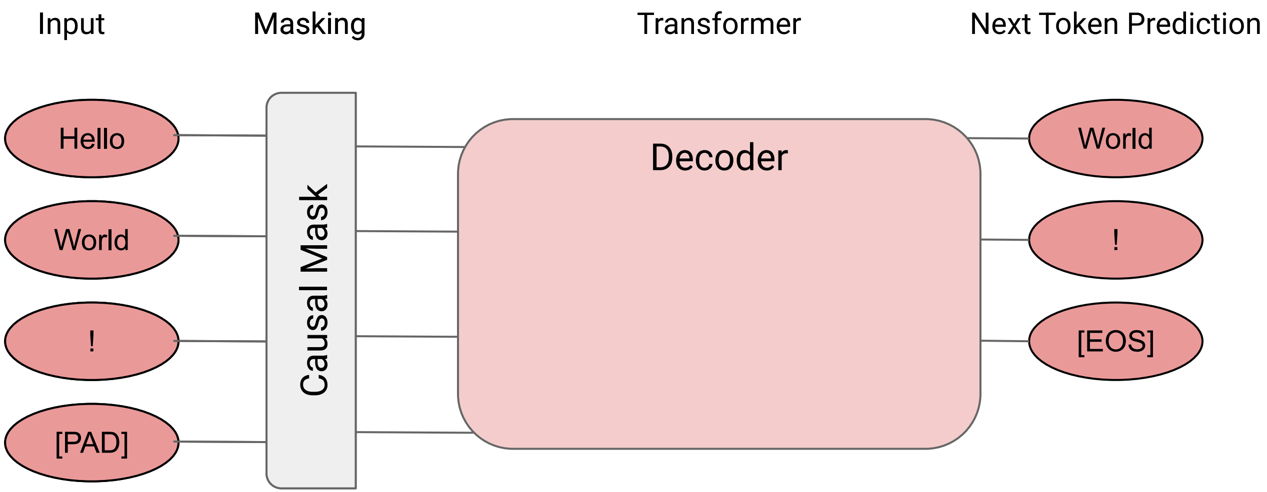
[Held et al. 2021](#)

Questions?



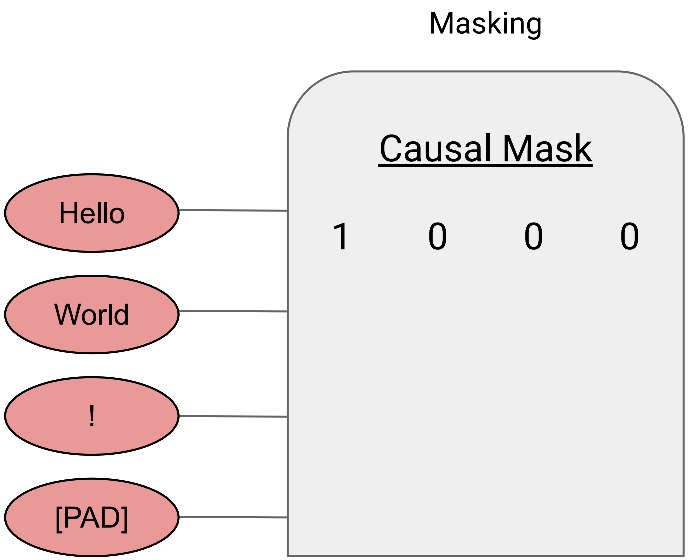
Masked Language Model

SSL | "How does GPT work?"



[Radford et al. 2019 (GPT-2)](#)

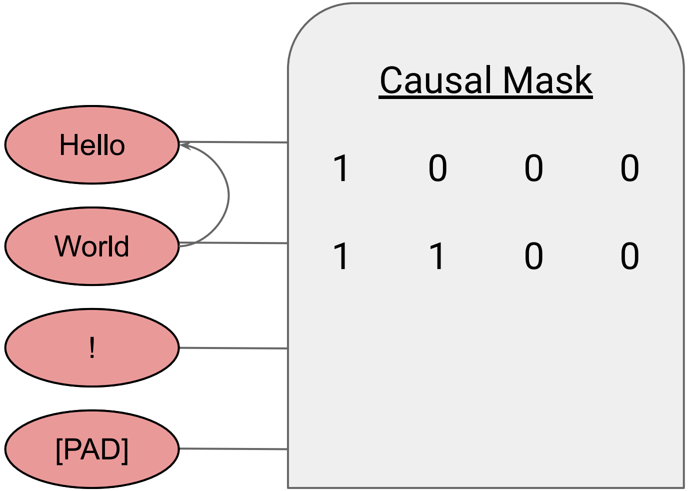
SSL | Autoregressive Language Modeling



[Radford et al. 2019 (GPT-2)](#)

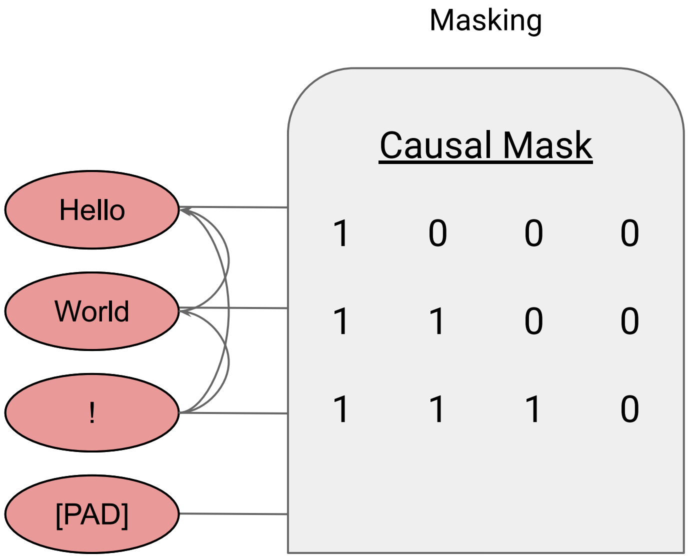
SSL | Autoregressive Language Modeling

Masking



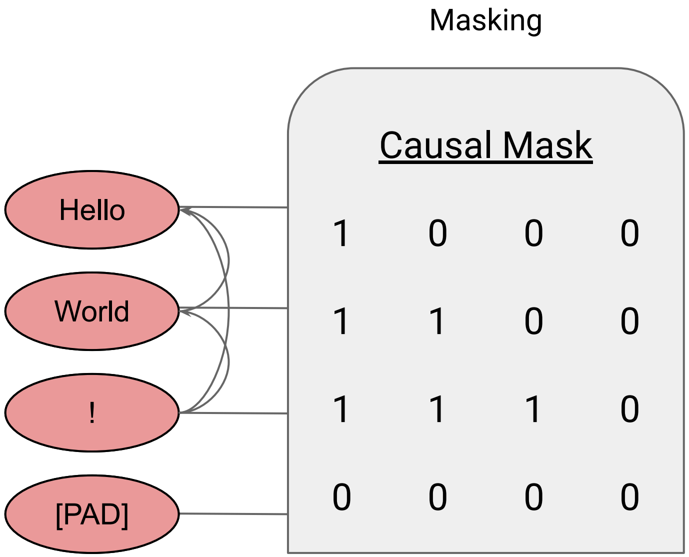
[Radford et al. 2019 (GPT-2)](#)

SSL | Autoregressive Language Modeling



[Radford et al. 2019 (GPT-2)]

SSL | Autoregressive Language Modeling



Masked Attention Again!

Similarities: E = (QXT / sqrt(DQ)) * MASK

Attention Matrix: A = softmax(E,dim=1)

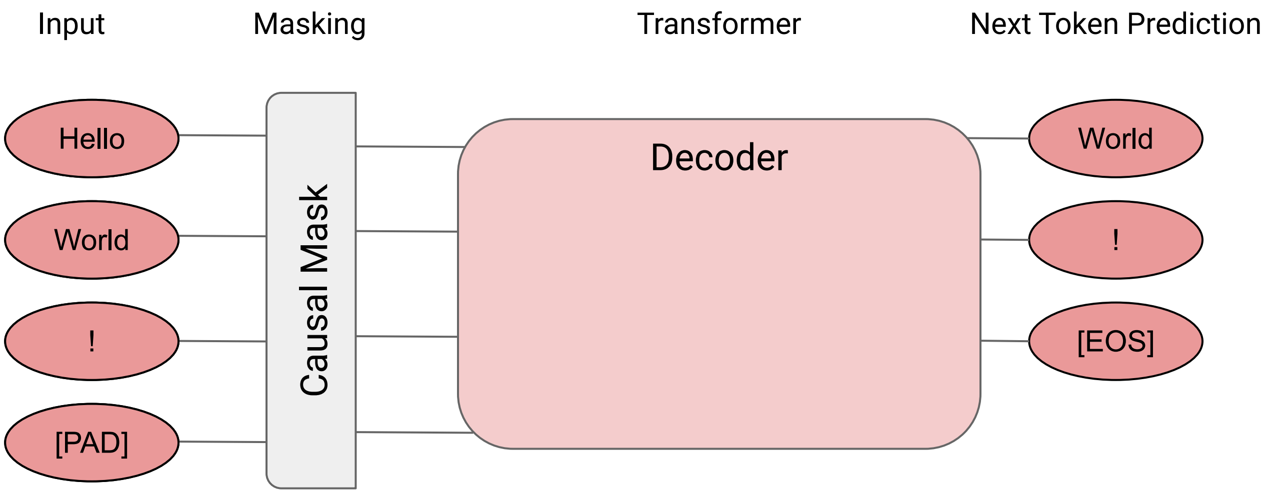
Output vectors: Y = AX

$Y_i = \sum_j A_{i,j} X$

Tokens only affected by preceding tokens

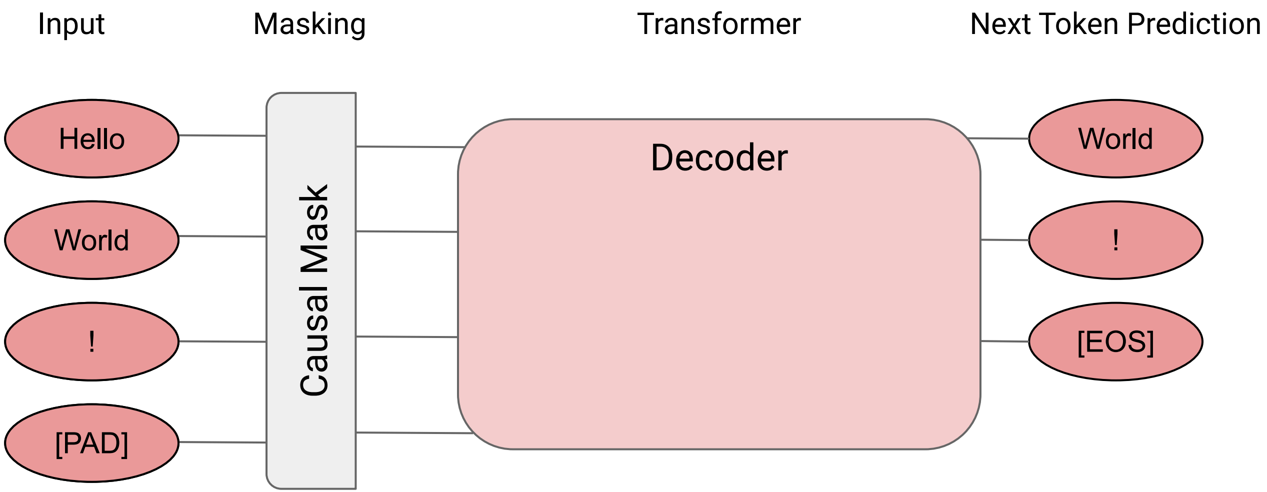
Radford et al. 2019 (GPT-2)

SSL | First successful GPT Model, Purely Autoregressive



[Radford et al. 2019 (GPT-2)](#)

SSL | First successful GPT Model, Purely Autoregressive



Optimize Negative Log Likelihood of Whole Sequence

```
loss = -(log(P(“World” | “Hello”) + log(P(“!” | “Hello World”) +
log(P(“[EOS]” | “Hello World!”))
```

[Radford et al. 2019 (GPT-2)](#)

Data | Increasing Token Count via Human Curation Heuristics

GPT-2 Corpus
All Reddit Outbound links with at least 3 karma

Size
~10 Billion Tokens

Quality
High quality text,
Broad Knowledge,
Improved Diversity

URL Domain	# Docs	% of Total Docs
bbc.co.uk	116K	1.50%
theguardian.com	115K	1.50%
washingtonpost.com	89K	1.20%
nytimes.com	88K	1.10%
reuters.com	79K	1.10%
huffingtonpost.com	72K	0.96%
cnn.com	70K	0.93%
cbc.ca	67K	0.89%
dailymail.co.uk	58K	0.77%
go.com	48K	0.63%

[Radford et al. 2019 (GPT-2)]

SSL | Architecture Comparison

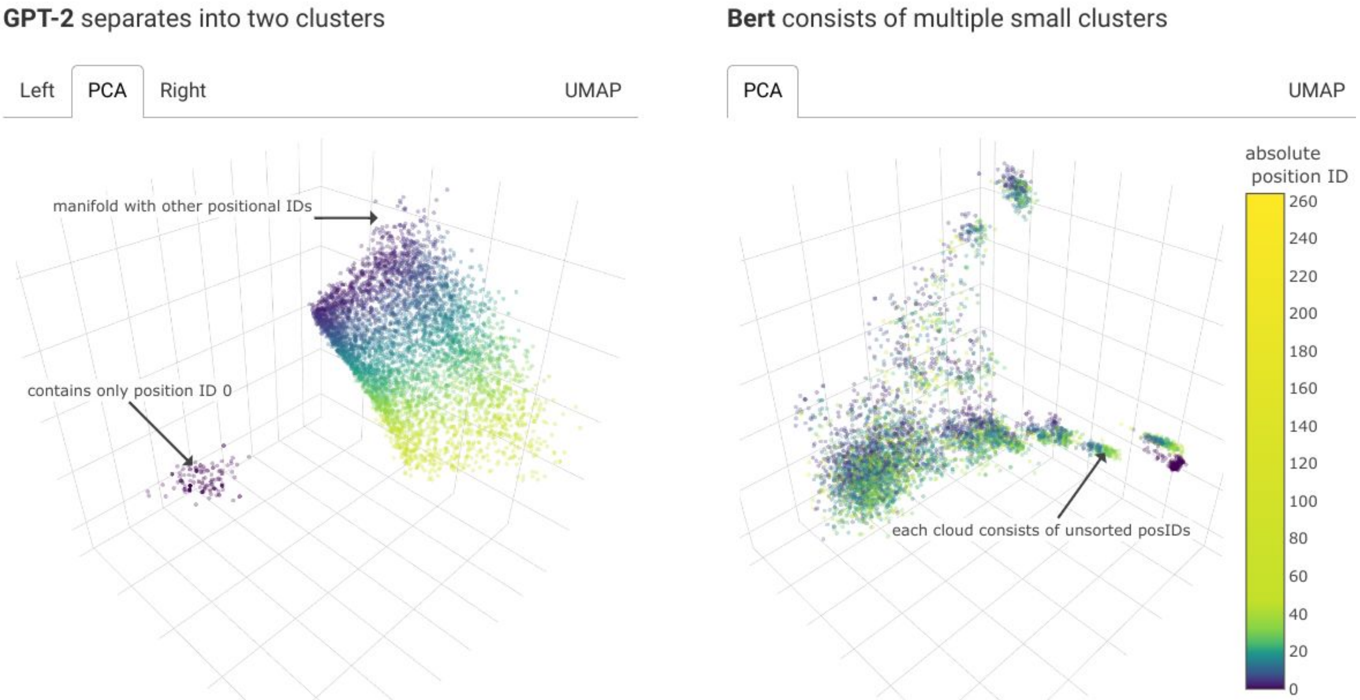
Ok, but what should I use?

SSL | Classification Comparison

Model	MNLI	CoLA	SST-2	MRPC	STS-B	QQP	QNLI	RTE	Avg
GPT-2-original	85.9/85.6	54.8	94.5	86.9/82.2	86.3/85.2	72.5/89.3	91.2	69.8	80.9
GPT-2-finetuned	85.8/85.5	40.9	94.5	87.0/81.0	85.6/84.3	71.4/88.5	91.5	69.0	78.8
RoBERTa-large	90.1/89.7	63.8	96.1	91.2/88.3	90.9/90.7	72.5/89.6	94.5	85.9	86.5

He et al. 2021

SSL | Pretrained Retrieval Comparison



https://bert-vs-gpt2.dbvis.de/

SSL | Generative Comparison

Encoders can’t generate!

SSL | Encoder-Only vs. Decoder-Only

Encoder

+ Retrieval
+ Classification
- No Generative Abilities

Decoder

+ Generative Abilities
- Retrieval
- Classification

Wang et al. 2022

SSL | Encoder-Only vs. Decoder-Only

Encoder

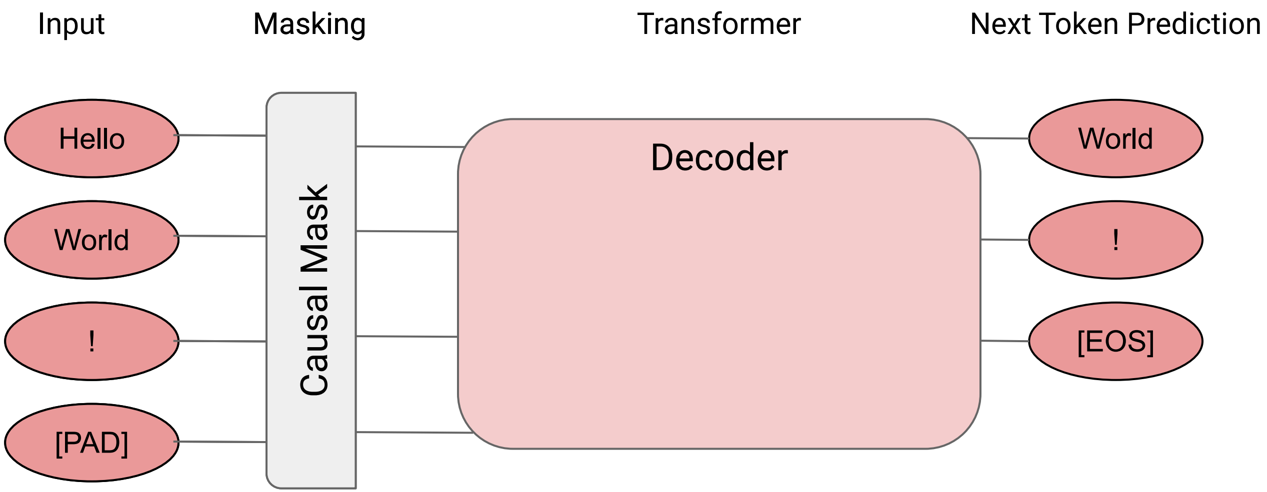
+ Retrieval
+ Classification
- No Generative Abilities

Decoder

+ $\boxed{\text{Generative Abilities}}$ — This is pretty essential!
- Retrieval
- Classification

Wang et al. 2022

Questions?



Autoregressive Language Model

SSL | Encoder-Only vs. Decoder-Only

Encoder

How to keep this?

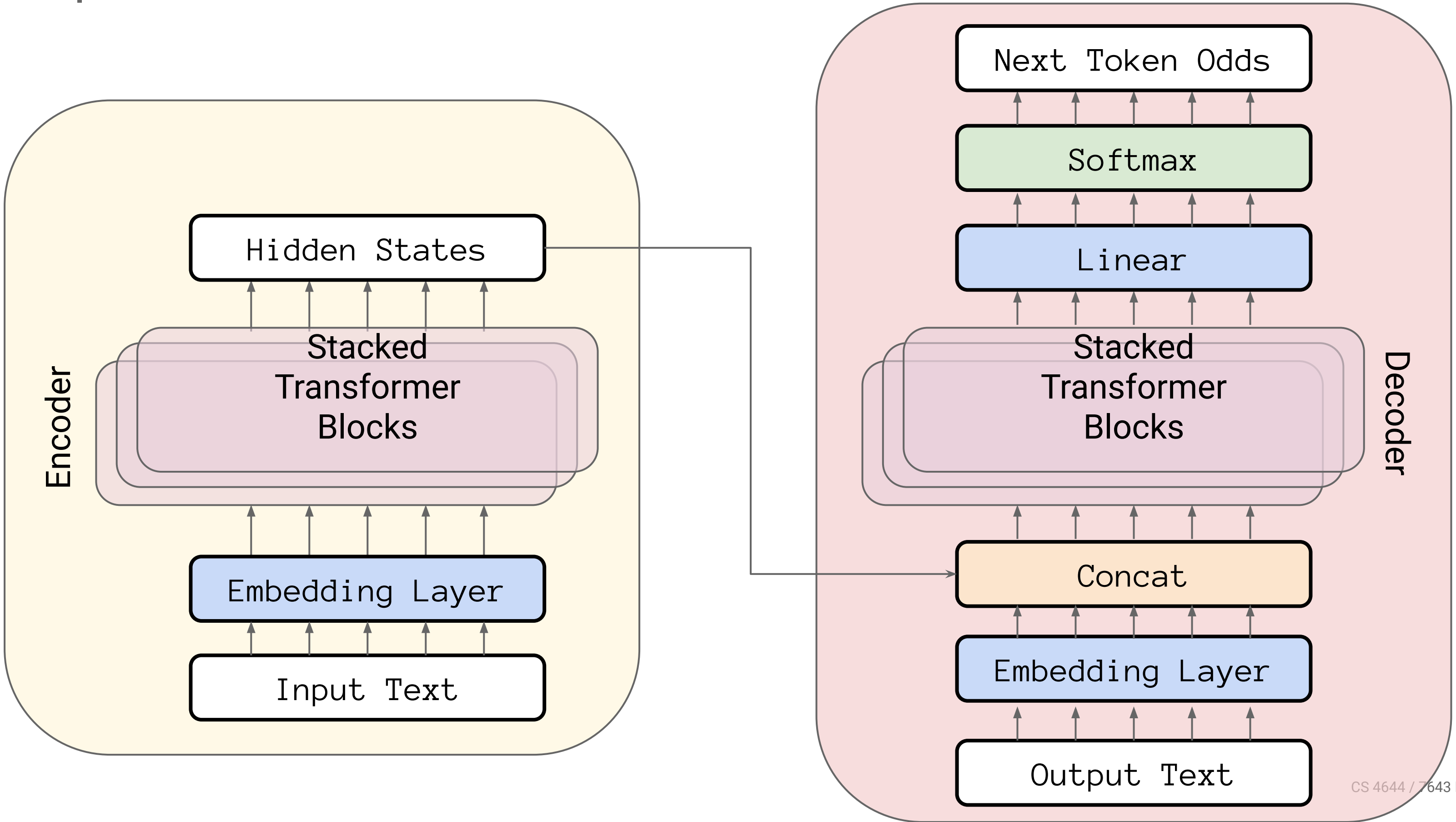
+ Retrieval
+ Classification
- No Generative Abilities

Decoder

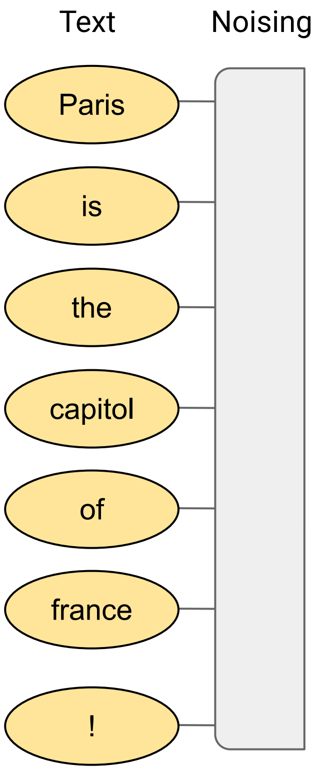
+ Generative Abilities
- Retrieval
- Classification

This is pretty essential!

SSL | Encoder-Decoder Returns

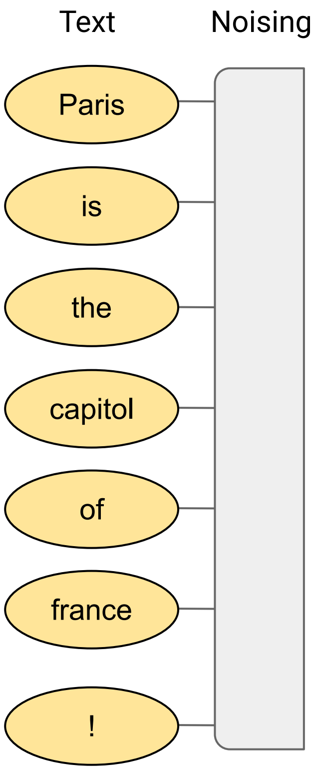


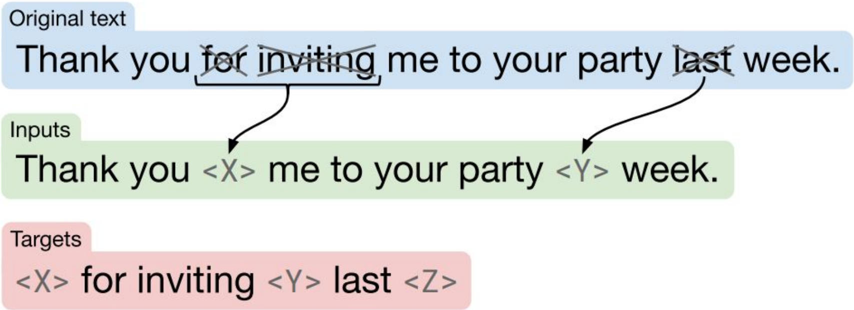
SSL | Universal Text-to-Text



Raffel et al. 2019

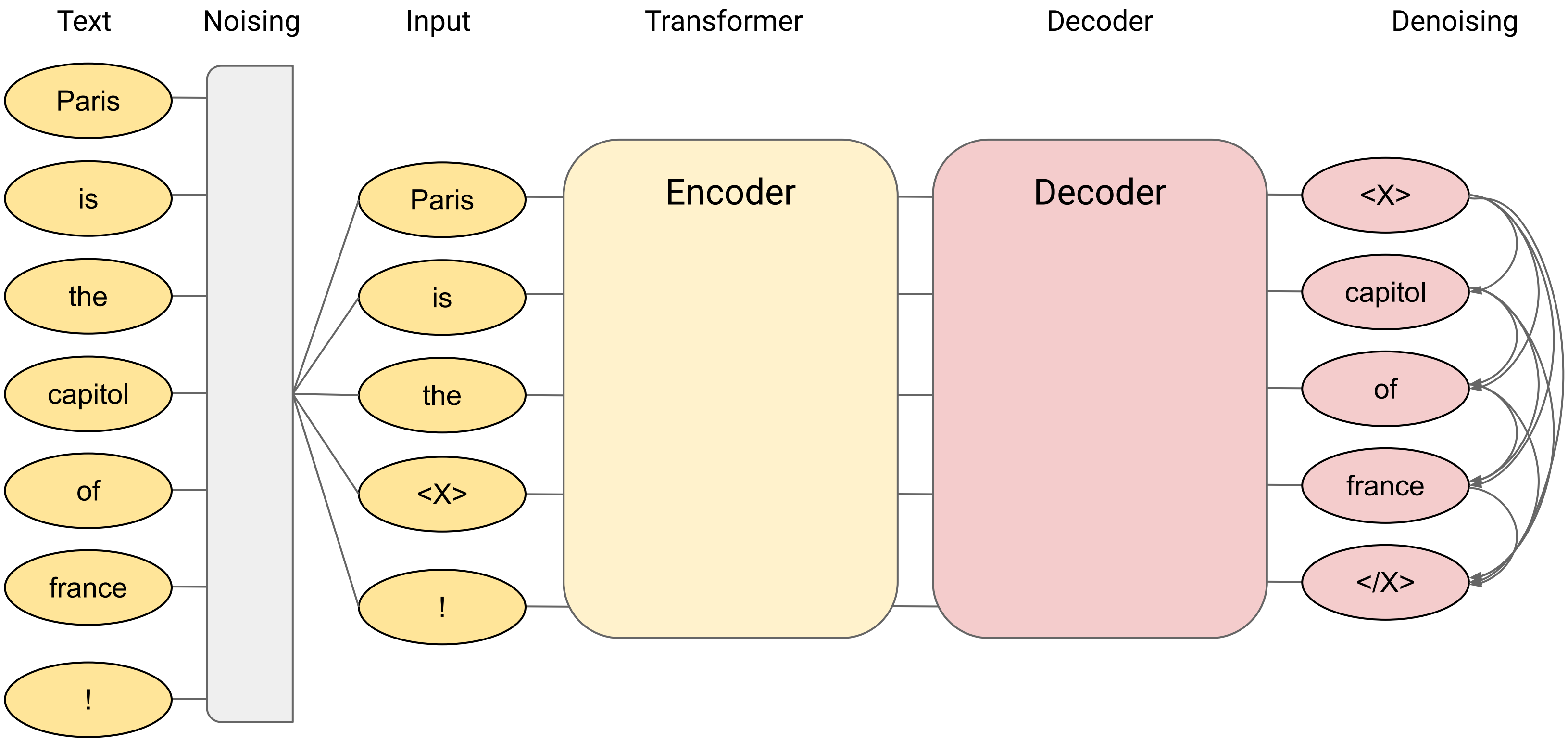
SSL | Universal Text-to-Text





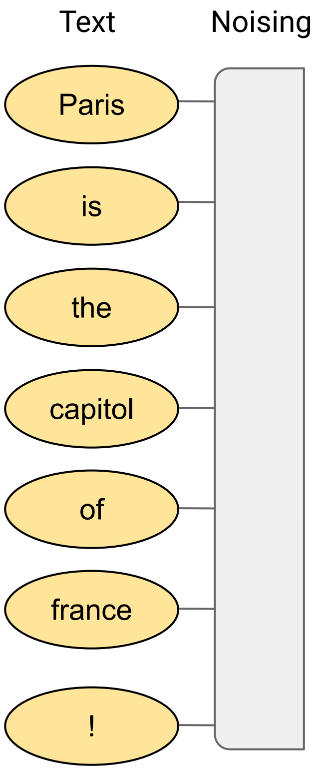
[Raffel et al. 2019]

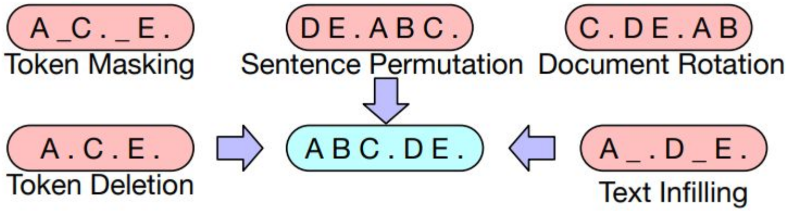
SSL | Universal Text-to-Text



[Raffel et al. 2019]

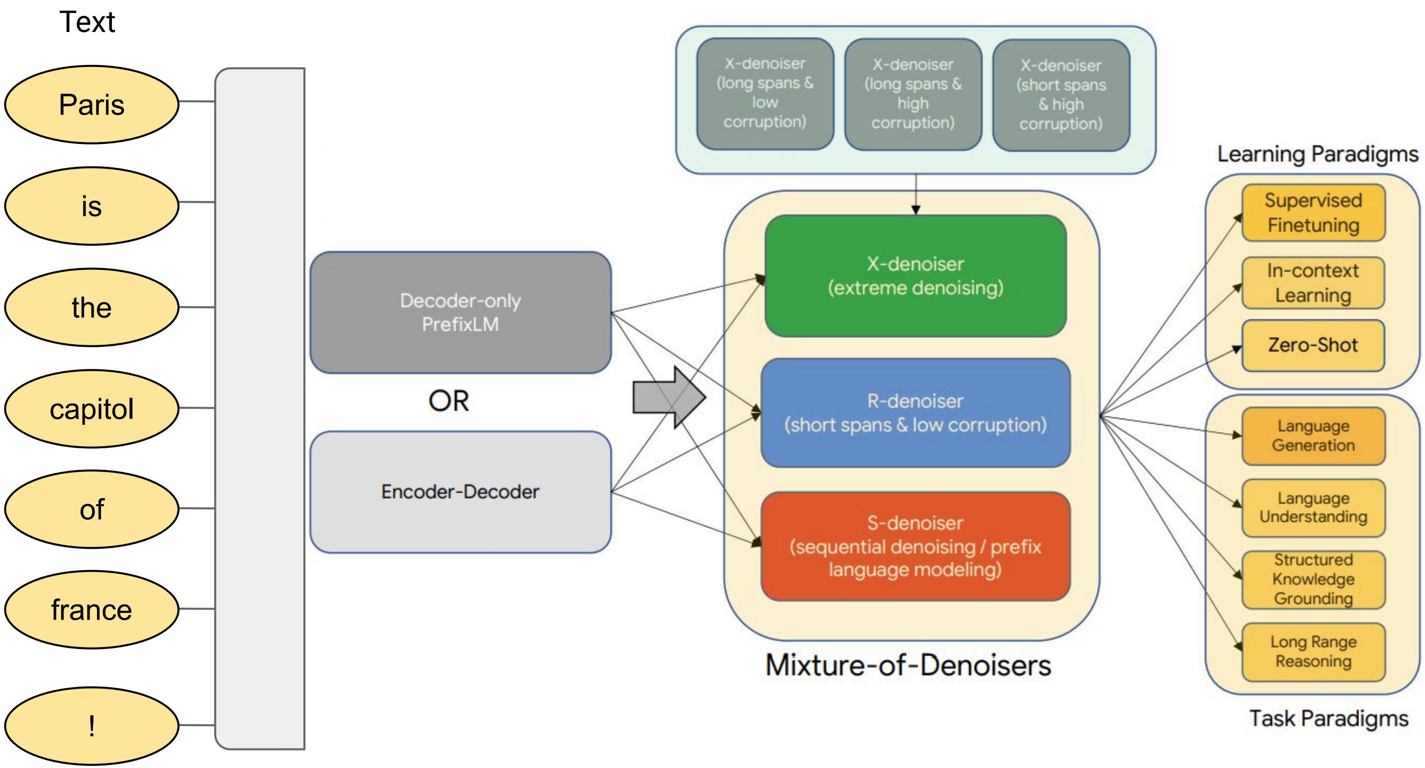
SSL | Universal Text-to-Text





[Lewis et al. 2020]

SSL | UL2 - Text-to-Text Pushed to Limits



[Tay et al. 2023]

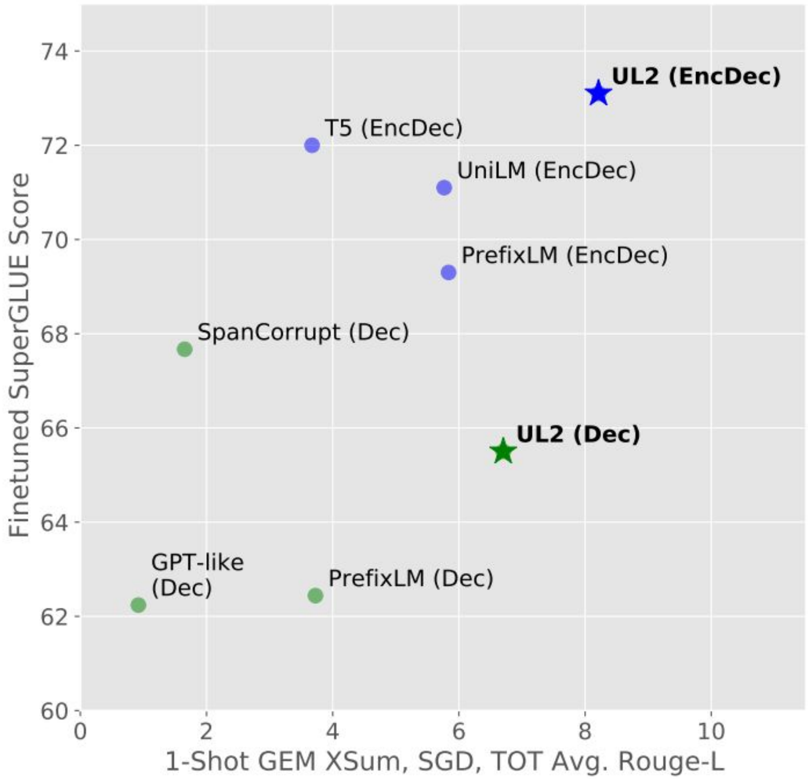
SSL | Universal Text-to-Text

Regardless of noise, Loss Function remains the same still!

Continue using Negative Log Likelihood

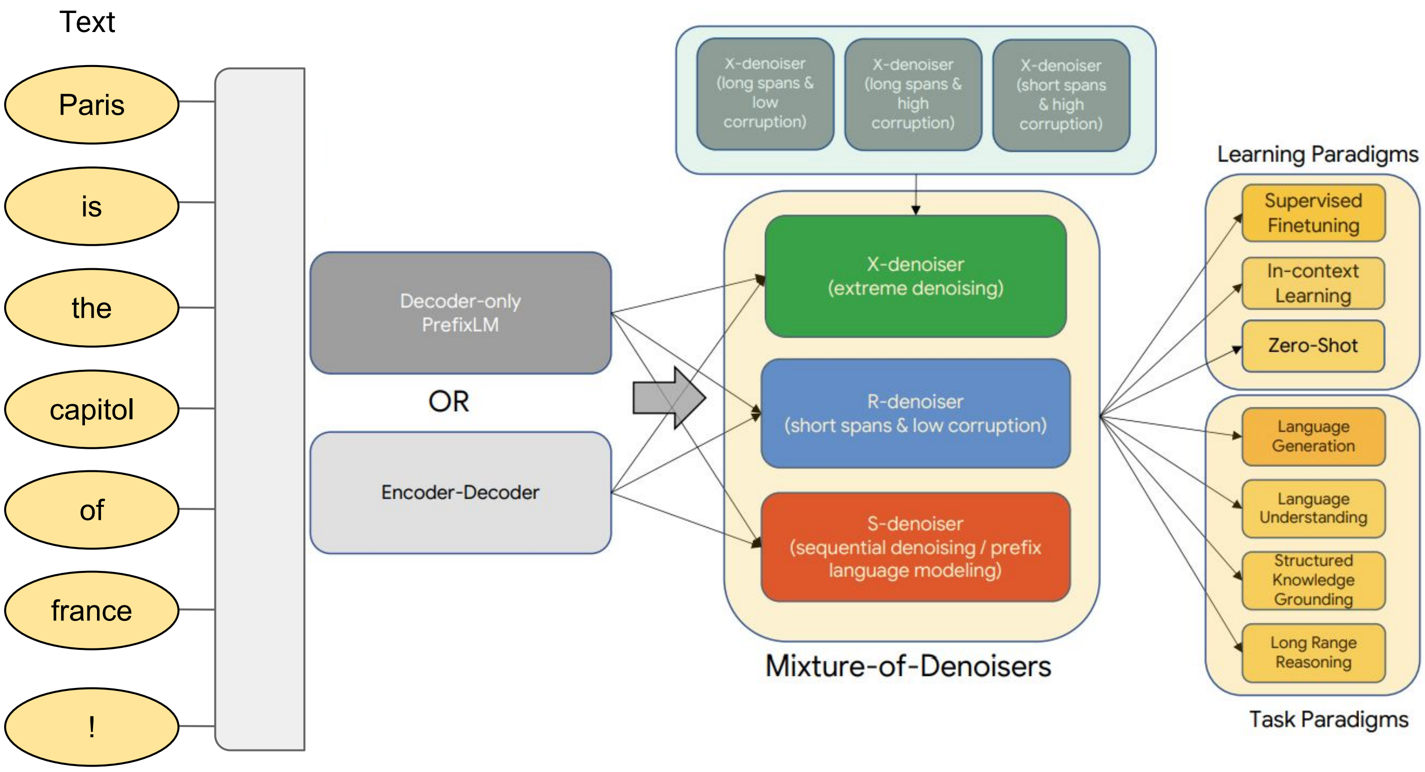
```
loss = -(log(P(Denoised Sequence | Noised Sequence))
```

SSL | Universal Text-to-Text Is Architecture Agnostic



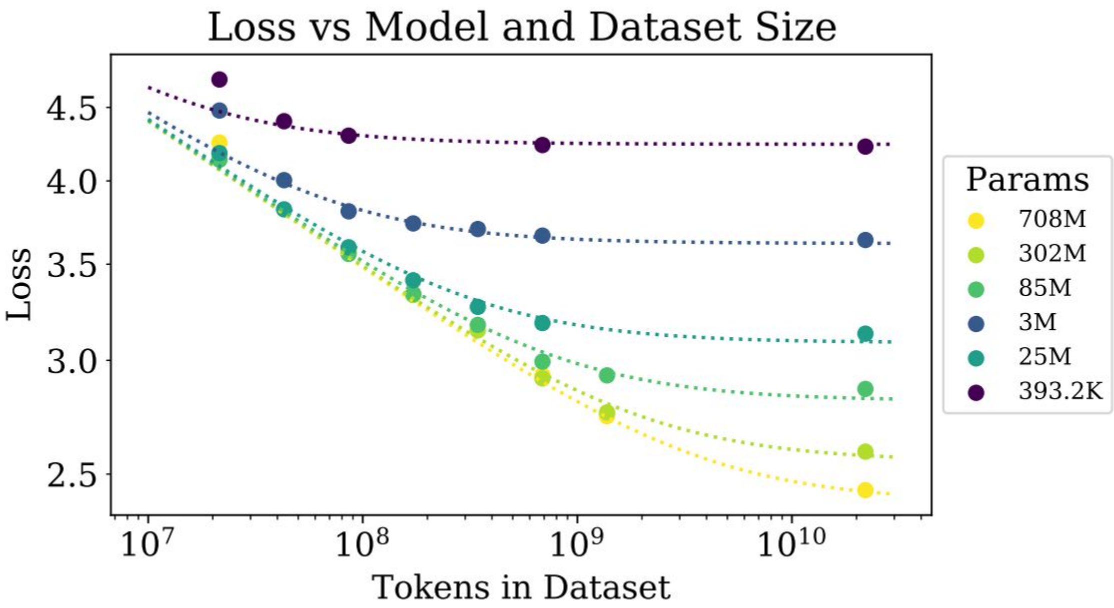
[Tay et al. 2023]

Questions?


Text
Paris
is
the
capitol
of
france
!
Decoder-only
PrefixLM
OR
Encoder-Decoder
X-denoiser
(long spans &
low
corruption)
X-denoiser
(long spans &
high
corruption)
X-denoiser
(short spans
& high
corruption)
X-denoiser
(extreme denoising)
R-denoiser
(short spans & low corruption)
S-denoiser
(sequential denoising / prefix
language modeling)
Mixture-of-Denoisers
Learning Paradigms
Supervised
Finetuning
In-context
Learning
Zero-Shot
Language
Generation
Language
Understanding
Structured
Knowledge
Grounding
Long Range
Reasoning
Task Paradigms

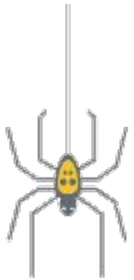
Data | Moving to truly Large Language Models

Today's LLMs are driven data and model scaling



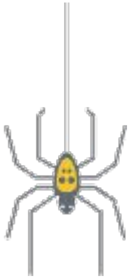
Kaplan et al. 2020

Data | Moving to truly Large Language Models



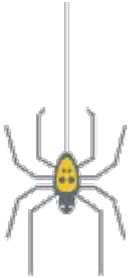
We could get a lot more data from CommonCrawl!

Data | Moving to truly Large Language Models



We could get a lot more data from CommonCrawl!
A lot of it is spam though…

Data | Moving to truly Large Language Models



We could get a lot more data from CommonCrawl!
A lot of it is spam though…
How do we get “useful” data?

Data | C4 - First Scaling of Data Via Common Crawl

T5 Corpus (AKA C4)
All Common Crawl Text Which Meets Heuristics

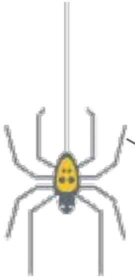
Size
~350 Billion Tokens

Quality
Varying quality text,
Broad Knowledge,
Improved Diversity

- We only retained lines that ended in a terminal punctuation mark (i.e. a period, exclamation mark, question mark, or end quotation mark).
- We discarded any page with fewer than 3 sentences and only retained lines that contained at least 5 words.
- We removed any page that contained any word on the "List of Dirty, Naughty, Obscene or Otherwise Bad Words".[6]
- Many of the scraped pages contained warnings stating that Javascript should be enabled so we removed any line with the word Javascript.
- Some pages had placeholder "lorem ipsum" text; we removed any page where the phrase "lorem ipsum" appeared.
- Some pages inadvertently contained code. Since the curly bracket "{" appears in many programming languages (such as Javascript, widely used on the web) but not in natural text, we removed any pages that contained a curly bracket.
- Since some of the scraped pages were sourced from Wikipedia and had citation markers (e.g. [1], [citation needed], etc.), we removed any such markers.
- Many pages had boilerplate policy notices, so we removed any lines containing the strings "terms of use", "privacy policy", "cookie policy", "uses cookies", "use of cookies", or "use cookies".
- To deduplicate the data set, we discarded all but one of any three-sentence span occurring more than once in the data set.

Raffel et al. 2019

Data | GPT-3 - Increased Scaling Via Curation

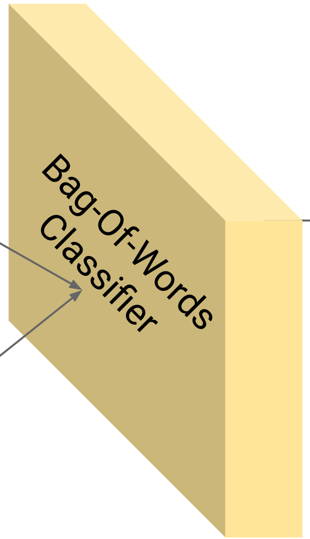


Low-Quality, High Volume

URL Domain	# Docs	% of Total Docs
bbc.co.uk	116K	1.50%
theguardian.com	115K	1.50%
washingtonpost.com	89K	1.20%
nytimes.com	88K	1.10%
reuters.com	79K	1.10%
huffingtonpost.com	72K	0.96%
cnn.com	70K	0.93%
cbc.ca	67K	0.89%
dailymail.co.uk	58K	0.77%
go.com	48K	0.63%

High Quality, Medium Volume

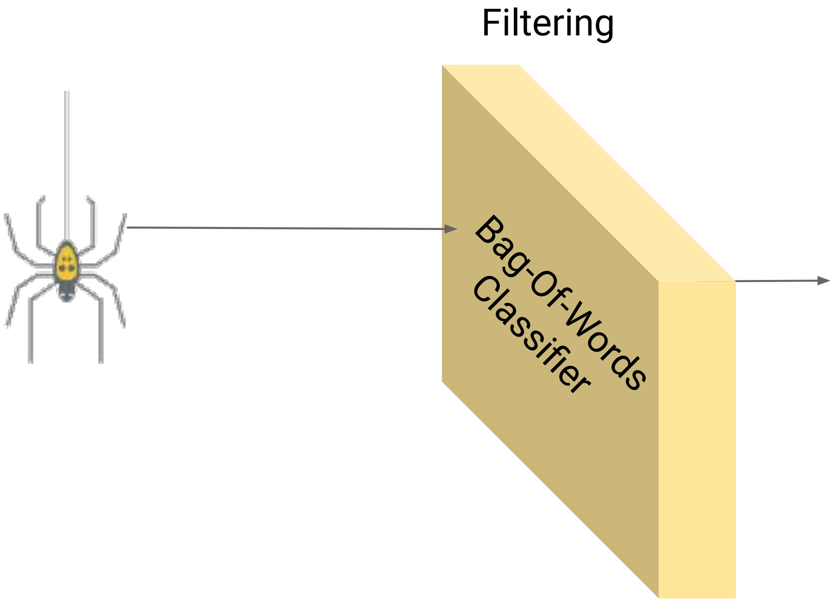
Training



Distinguish High and Low Quality

Brown et al. 2020

Data | GPT-3 - Increased Scaling Via Curation



Keep "False" Positives

"False" positive ~= High Quality

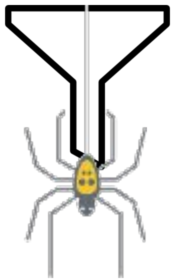
Brown et al. 2020

Data | GPT-2 to Original GPT-3 was mostly data scaling

GPT-3 Corpus
Common-Crawl Filtered using
GPT-2 Training Data

Size
~400 Billion Tokens

Quality
High-ish quality text,
Broad Knowledge,
Web-scale Diversity



Brown et al. 2020

Data | Recent Open Source models focus heavily on data scaling

Llama 1 Corpus

Size
~1.4 Trillion Tokens

Quality
Varying quality text,
Broad Knowledge,
Web-scale Diversity,
Includes Code!

Dataset	Sampling prop.	Epochs	Disk size
CommonCrawl	67.0%	1.10	3.3 TB
C4	15.0%	1.06	783 GB
Github	4.5%	0.64	328 GB
Wikipedia	4.5%	2.45	83 GB
Books	4.5%	2.23	85 GB
ArXiv	2.5%	1.06	92 GB
StackExchange	2.0%	1.03	78 GB

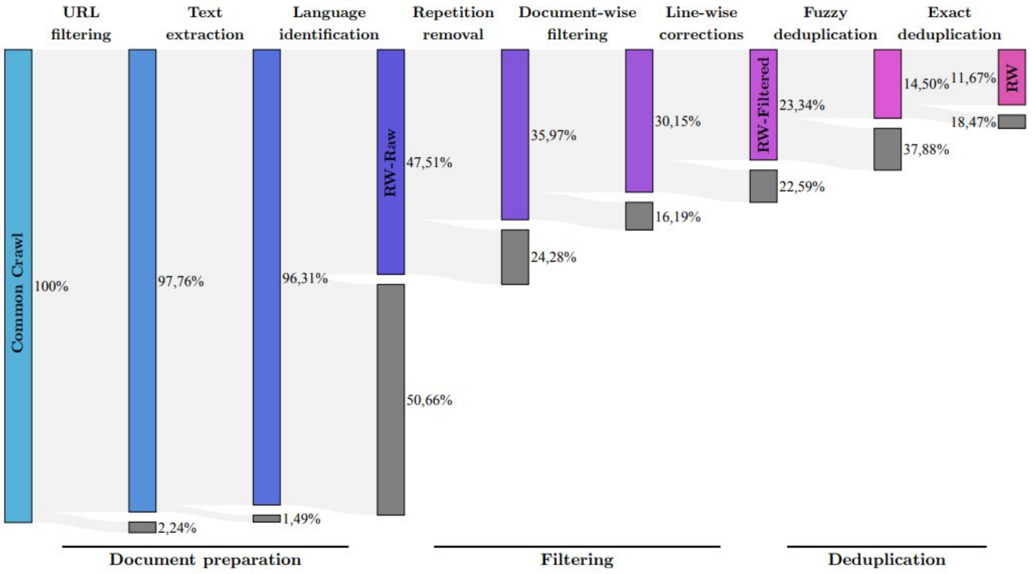
Touvron et al. 2023

Data | Recent Open Source models focus heavily on data scaling

<u>Falcon Refined Web Corpus</u>

<u>Size</u>
5 Trillion Tokens

<u>Quality</u>
Varying quality text,
Broad Knowledge,
Web-scale Diversity,
Includes Code



[Brown et al. 2020]

Data | Data Mixture has become the biggest "secret"

Llama 2 Corpus

Size
> 2 Trillion Tokens

Quality
Minimal details known

Touvron et al. 2023 (b)

PALM-2 Corpus

Size
> 3.6 Trillion Tokens

Quality
No details known

Anil et al. 2023

GPT-4 Corpus

Size
Unknown (Est. 11T Tokens)

Quality
No details known

OpenAI 2023

Questions?

Llama 2 Corpus

Size
> 2 Trillion Tokens

Quality
Minimal details known

Touvron et al. 2023 (b)

PALM-2 Corpus

Size
> 3.6 Trillion Tokens

Quality
No details known

Anil et al. 202320

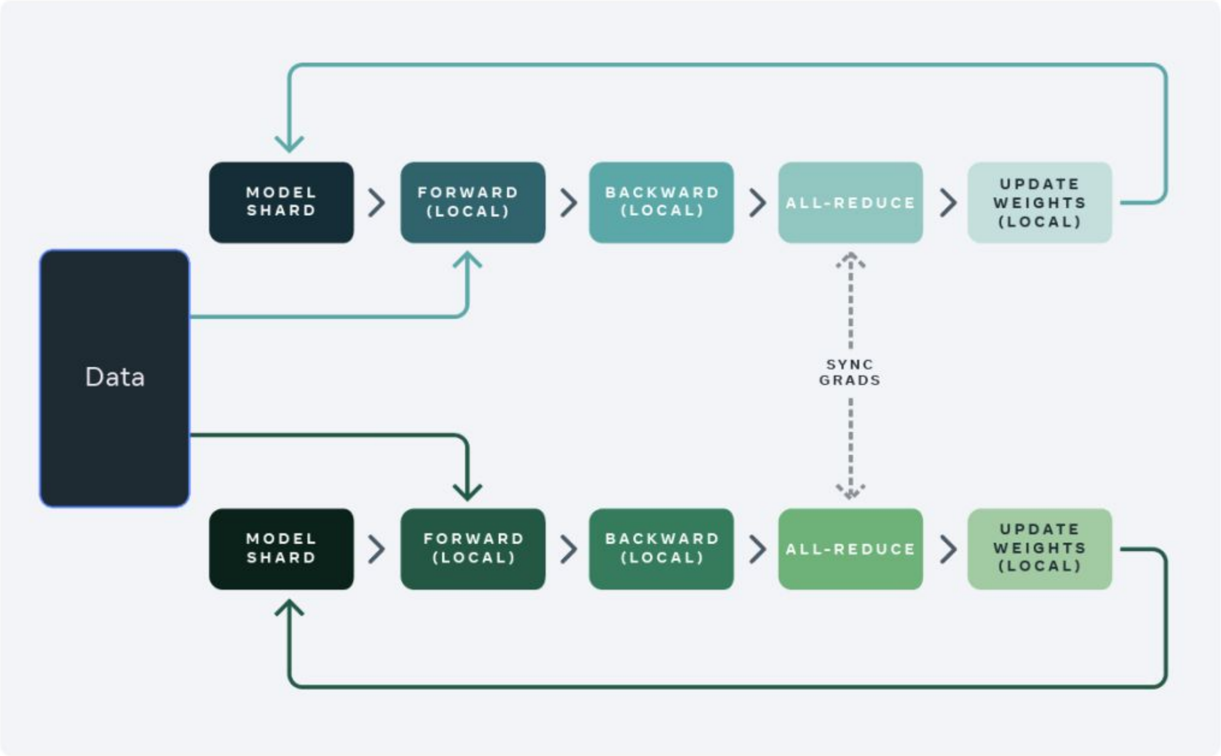
GPT-4 Corpus

Size
Unknown (Est. 11T Tokens)

Quality
No details known

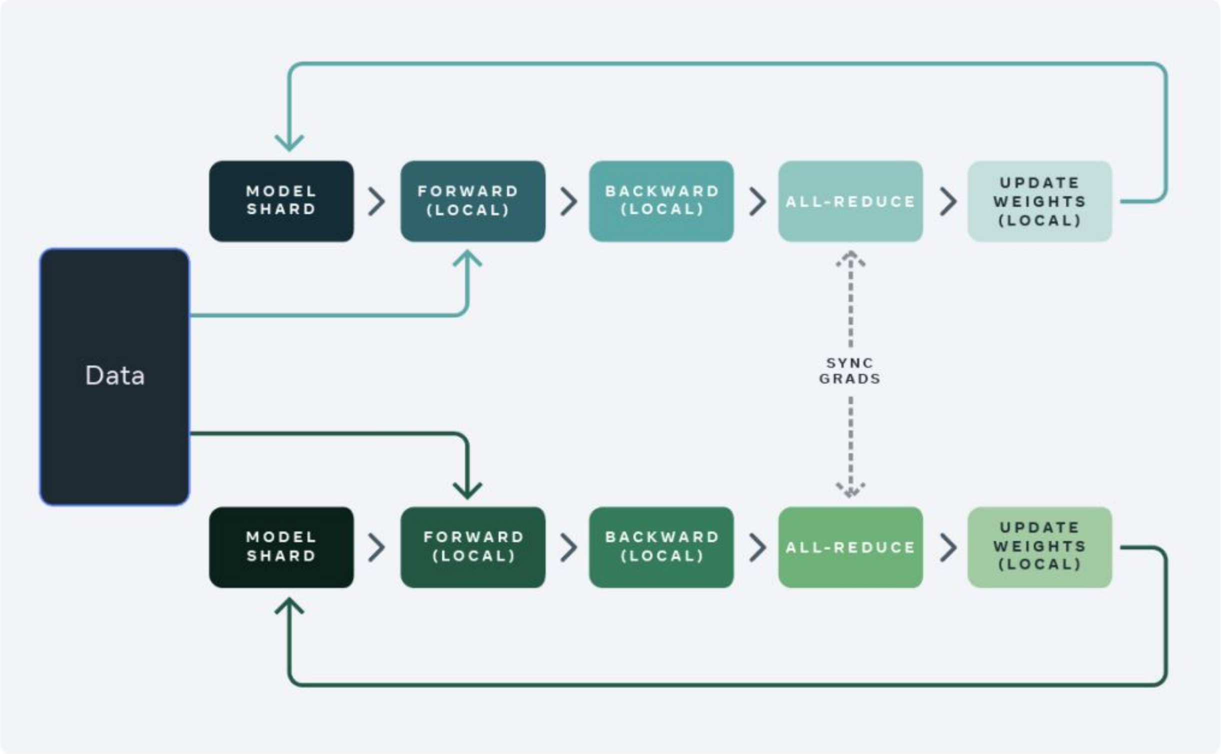
OpenAI 2023

Scaling Parameters | Data Parallel Training



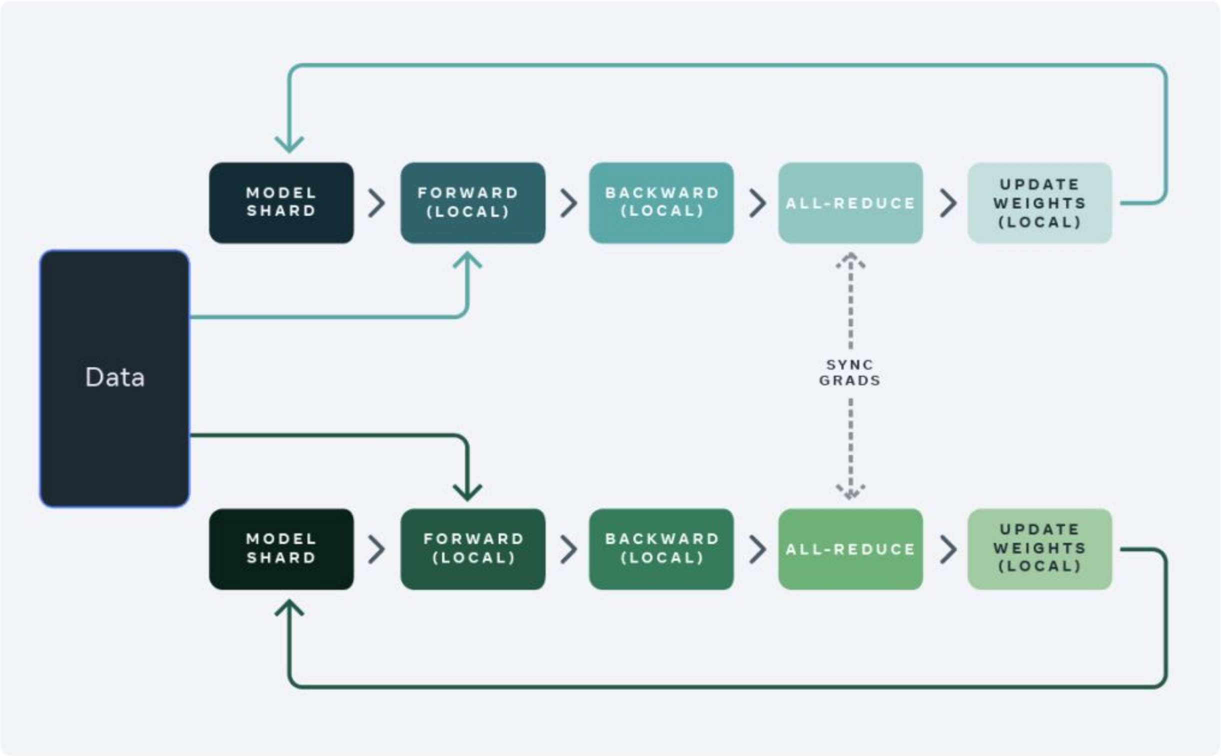
https://engineering.fb.com/2021/07/15/open-source/fsdp/

Scaling Parameters | Data Parallel Training



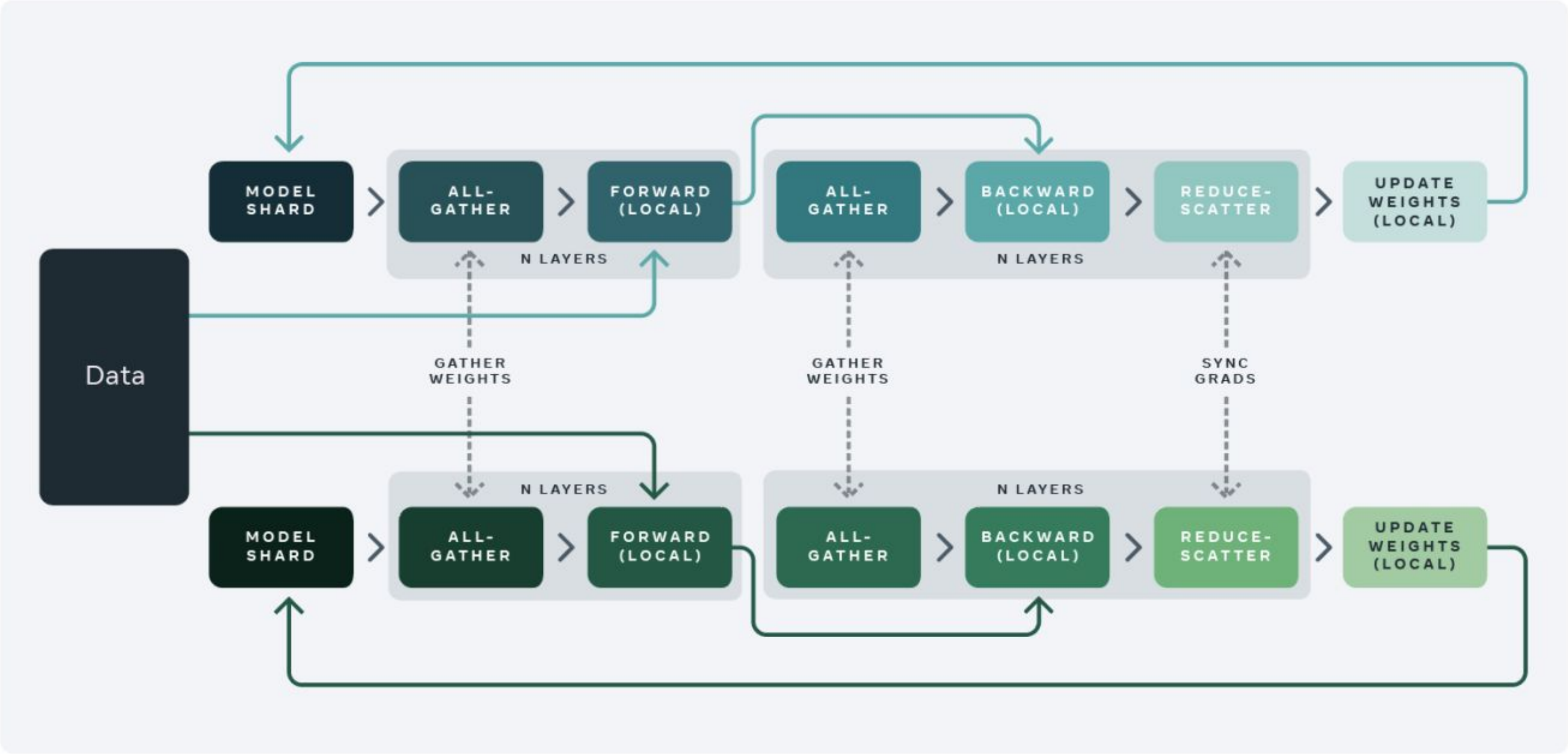
Total memory increases linearly with shards

Scaling Parameters | Data Parallel Training



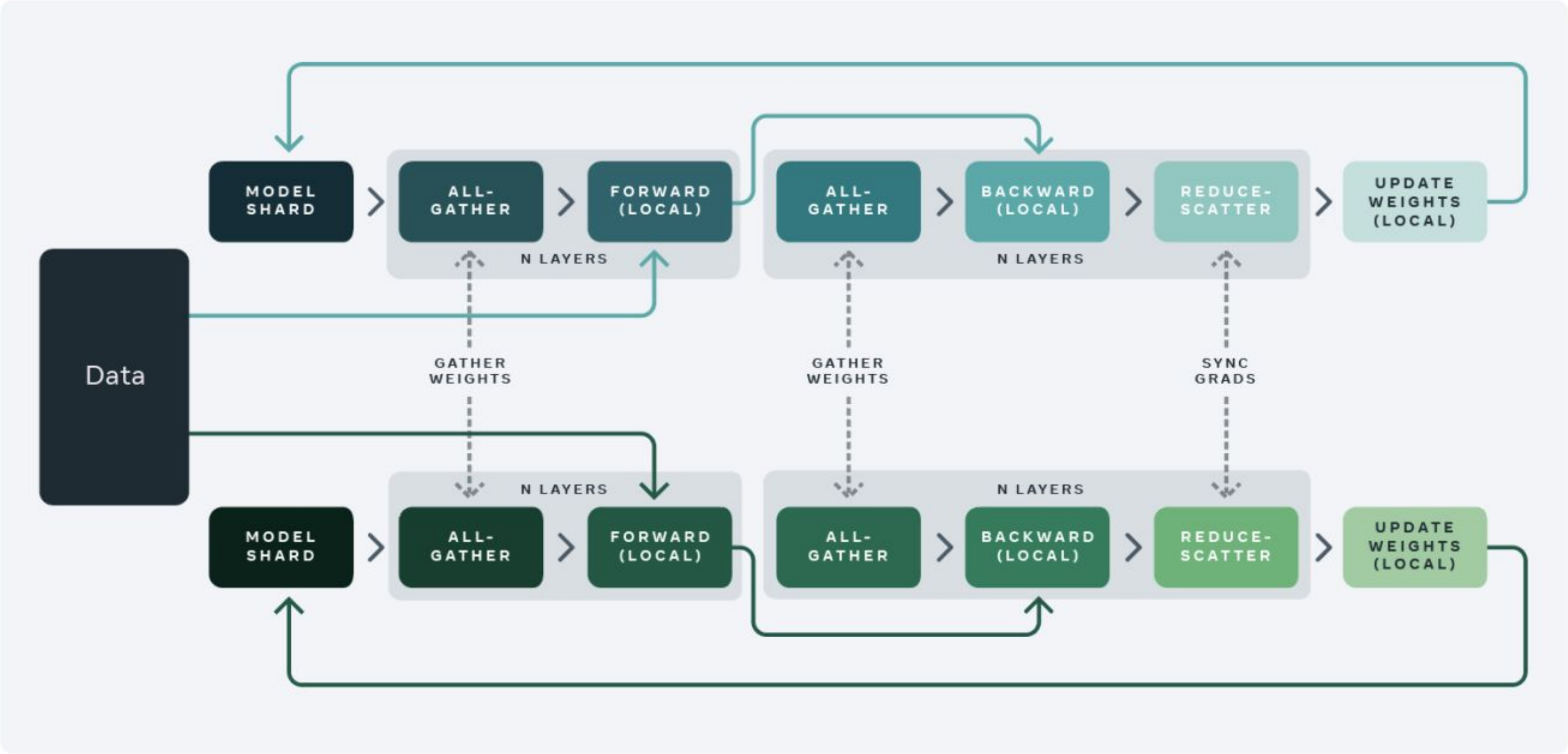
Max memory constrains model size

Scaling Parameters | *Fully* Sharded Data Parallel Training



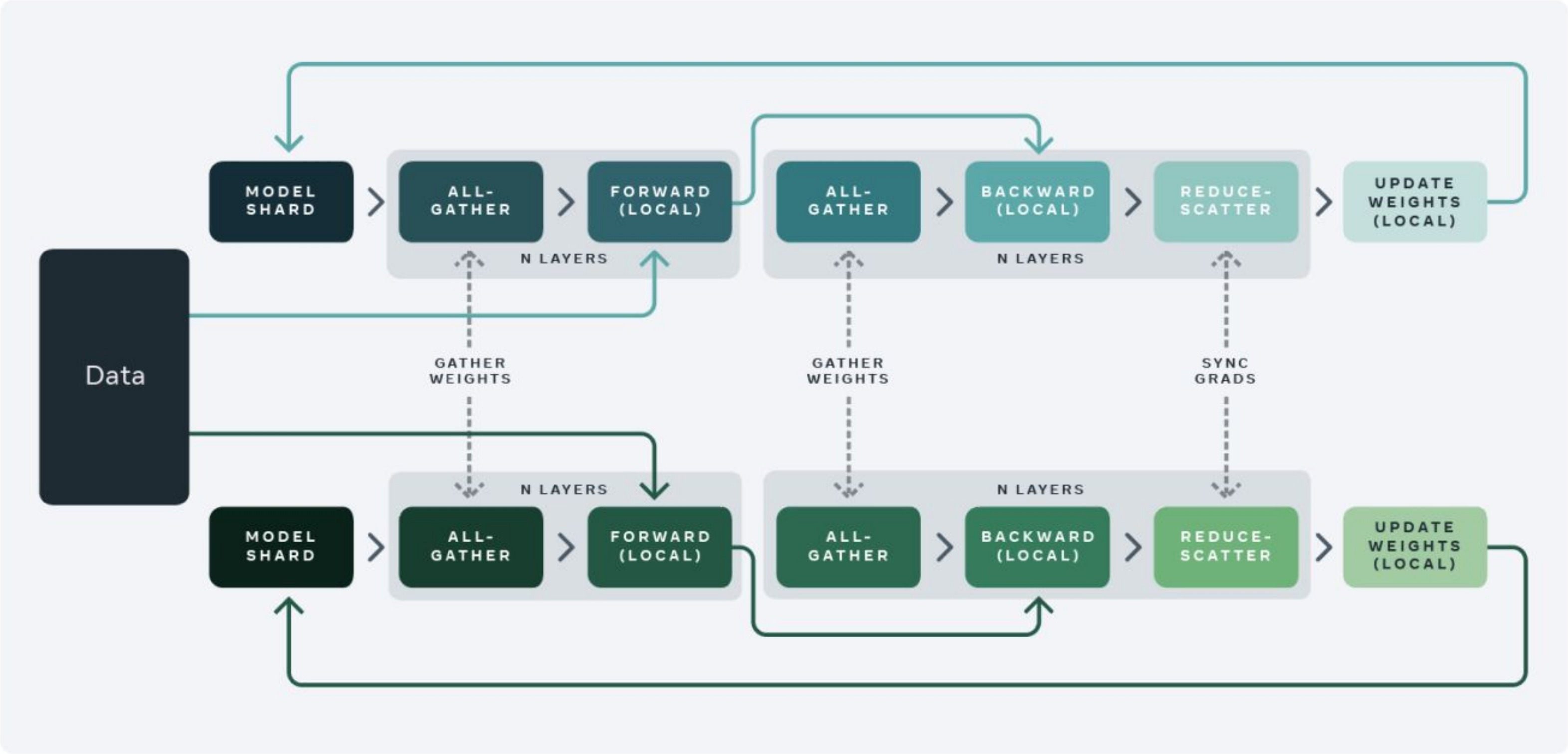
https://engineering.fb.com/2021/07/15/open-source/fsdp/

Scaling Parameters | *Fully* Sharded Data Parallel Training



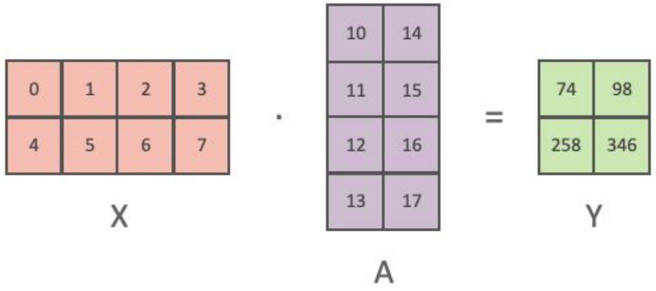
Total memory is constant

Scaling Parameters | *Fully* Sharded Data Parallel Training



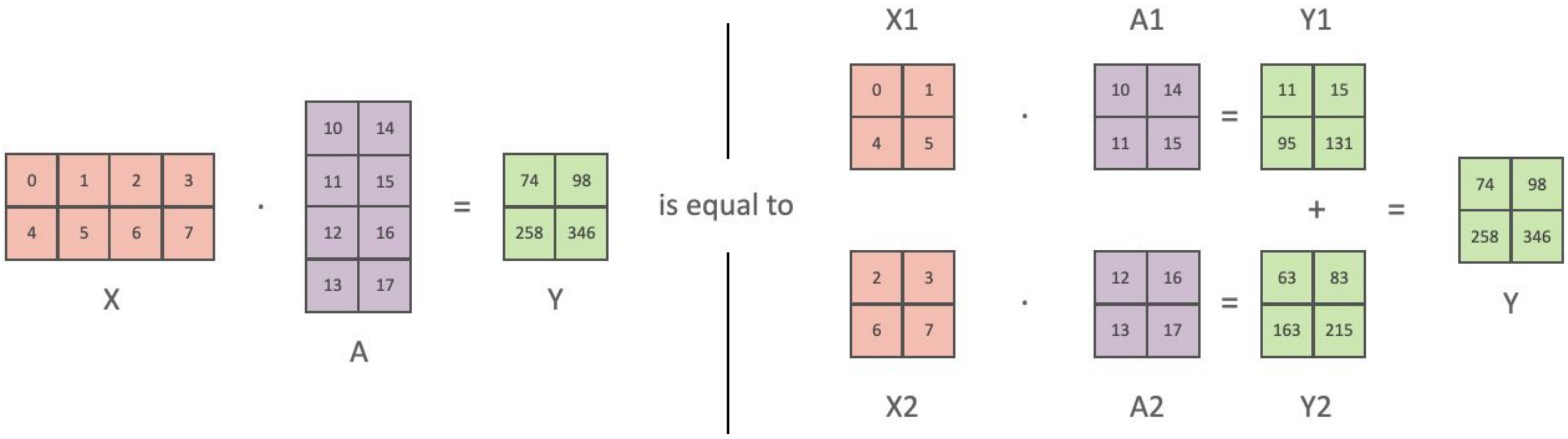
Max single GPU memory constrains layer size

Scaling Parameters | Tensor Parallel Training



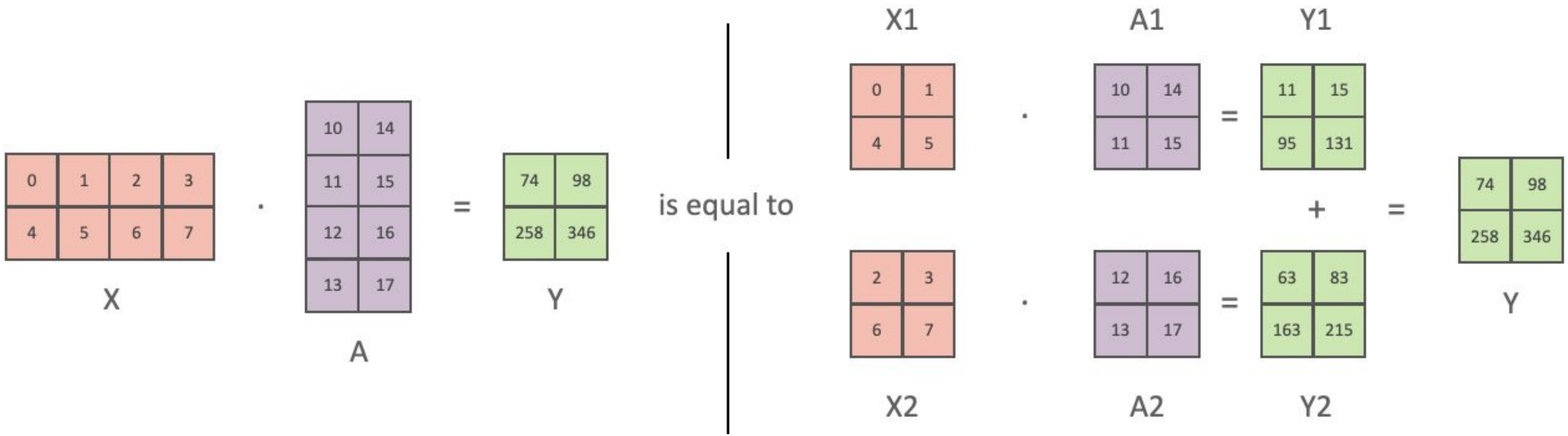
https://huggingface.co/docs/transformers/v4.15.0/parallelism#tensor-parallelism

Scaling Parameters | Tensor Parallel Training



https://huggingface.co/docs/transformers/v4.15.0/parallelism#tensor-parallelism

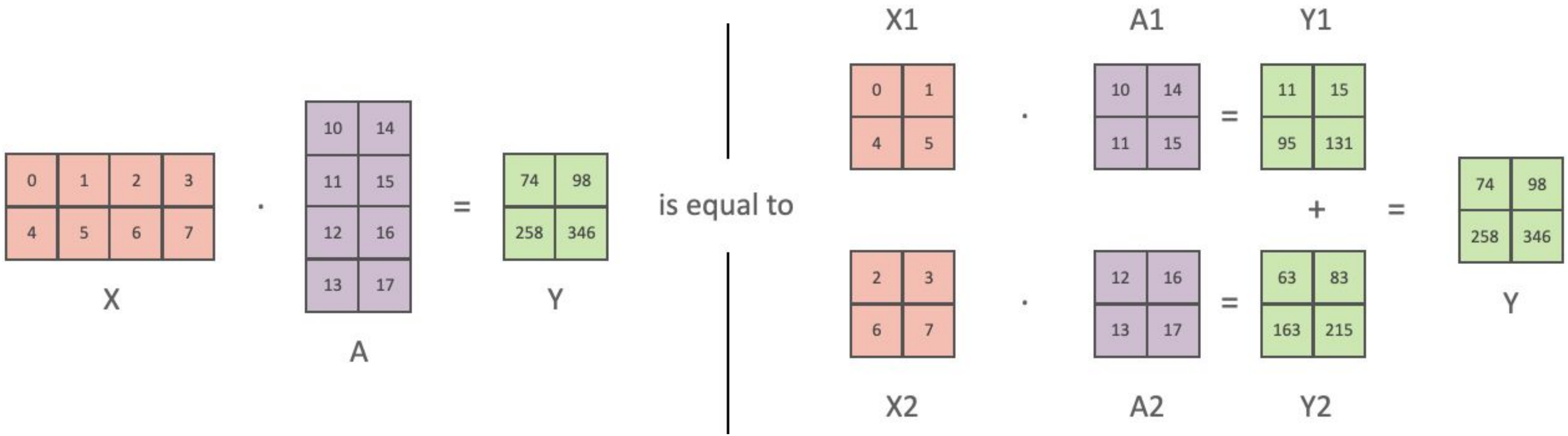
Scaling Parameters | Tensor Parallel Training



Don’t need to sync gradients!

https://huggingface.co/docs/transformers/v4.15.0/parallelism#tensor-parallelism

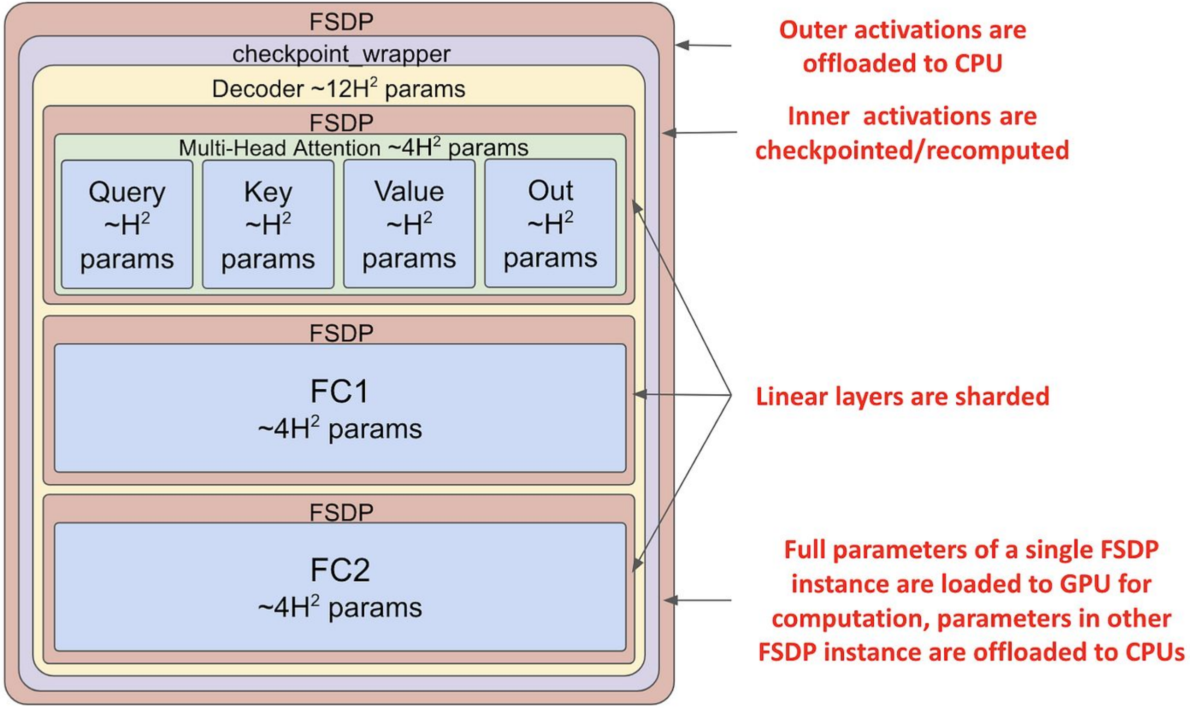
Scaling Parameters | Tensor Parallel Training



Don't need to sync gradients!
Max GPU memory constrains a layer shard

https://huggingface.co/docs/transformers/v4.15.0/parallelism#tensor-parallelism

Scaling Parameters | FSDP + TP = ~Limitless Scaling



[1 Trillion Parameter Model with Tensor Parallelism and FSDP]

Final Questions?

Fill out my anonymous feedback form