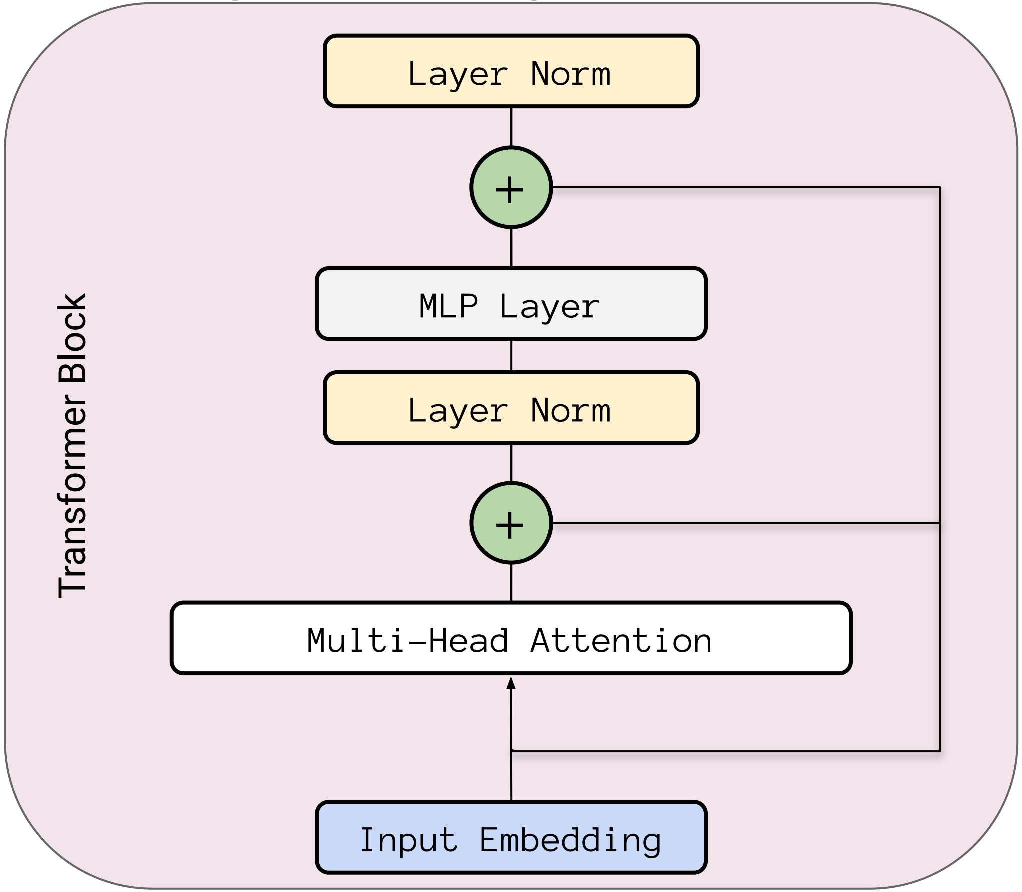
Training Large Language Models

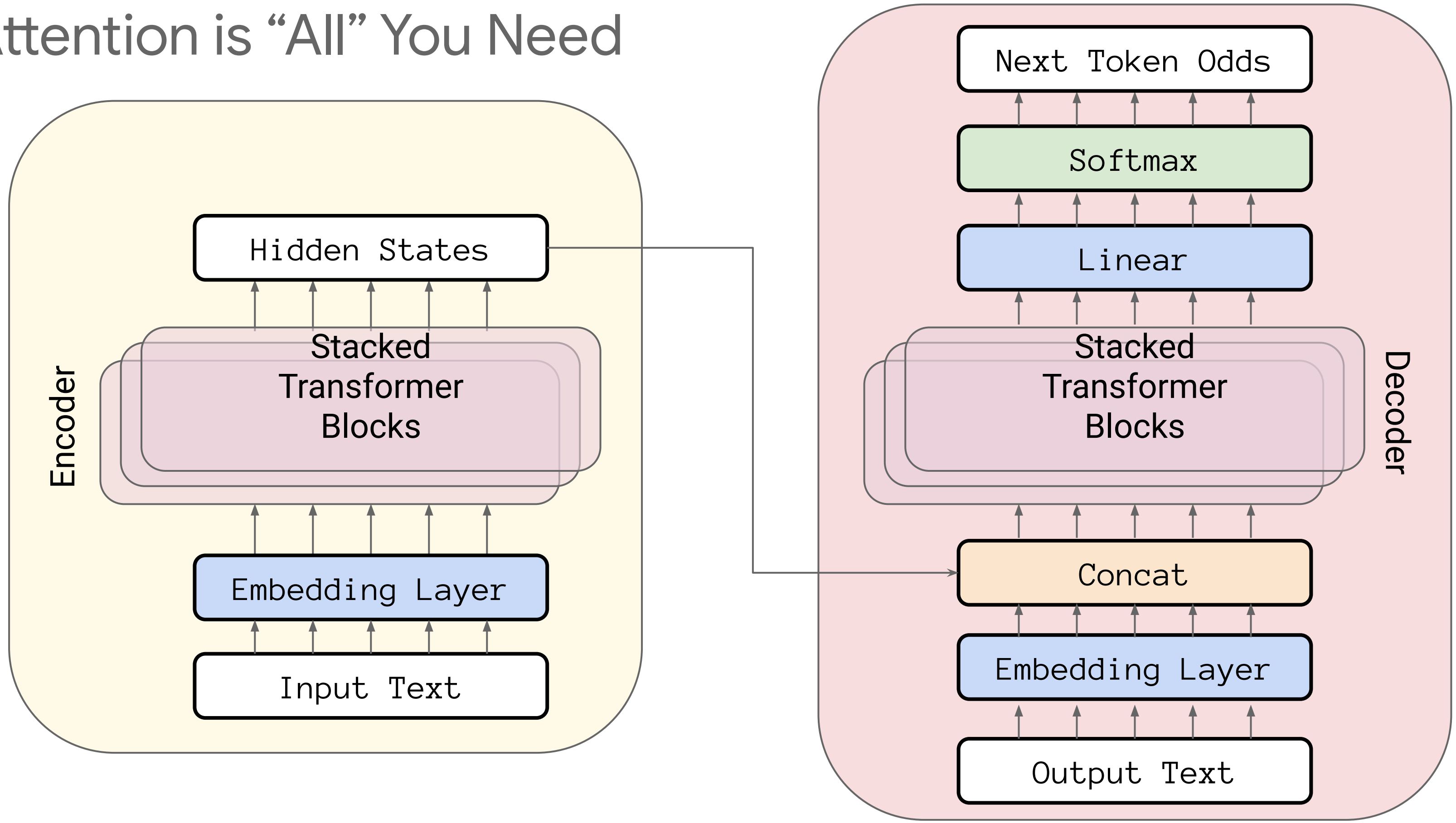
CS 4644 / 7643: Deep Learning

William Held
School of Interactive Computing
Georgia Institute of Technology

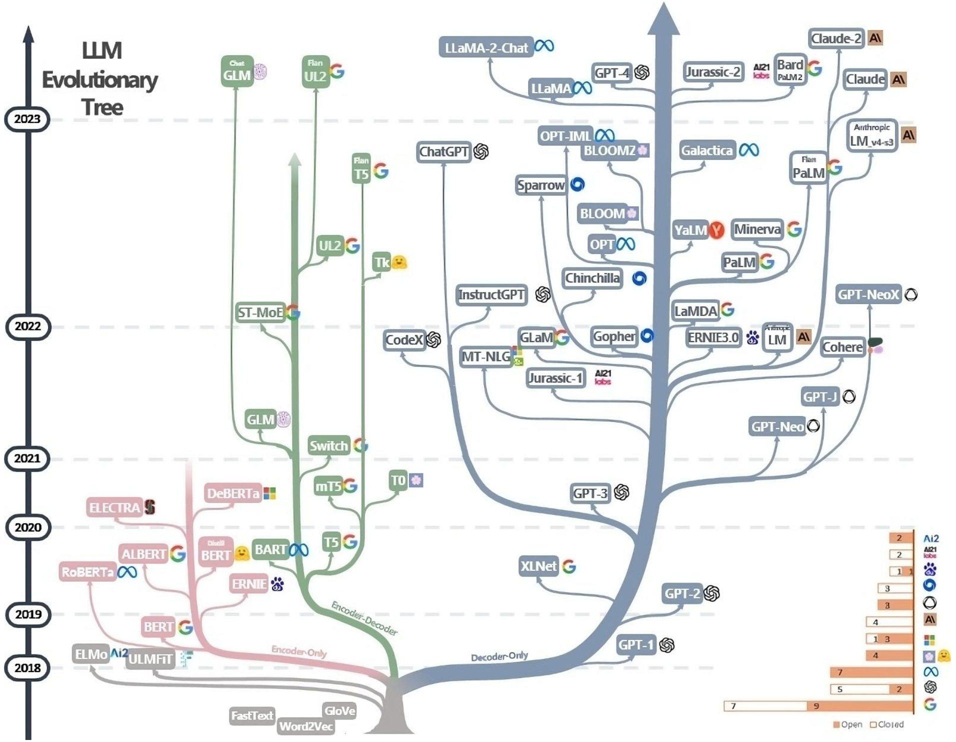
Transformer Lecture Speed Recap: The Transformer Block



Attention is "All" You Need



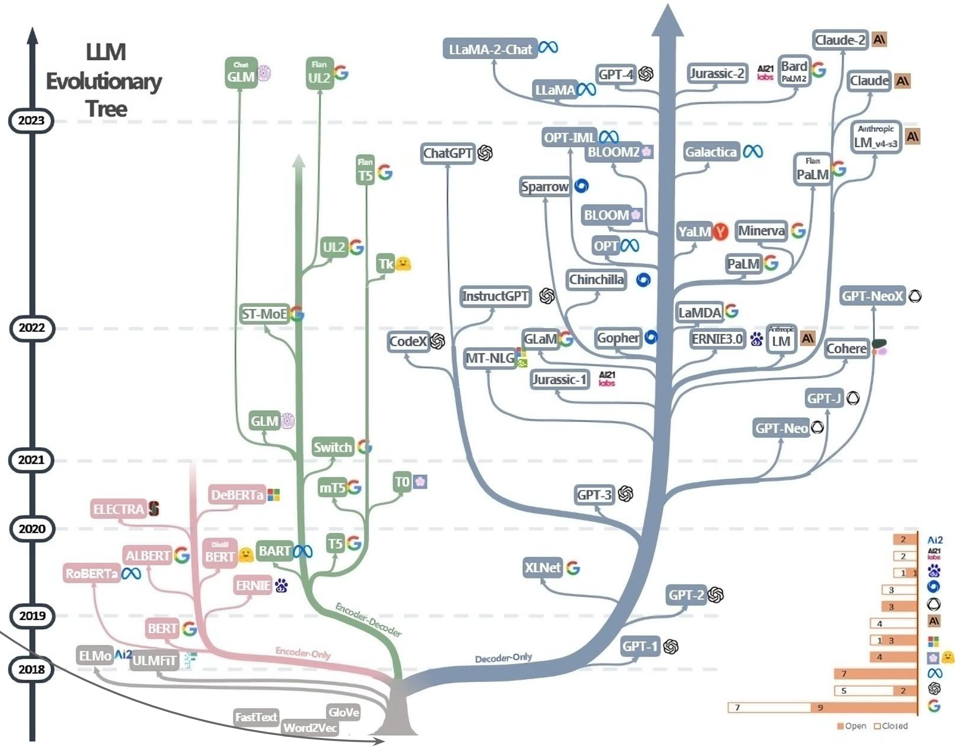
How do we go from purpose driven models to LLMs?



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

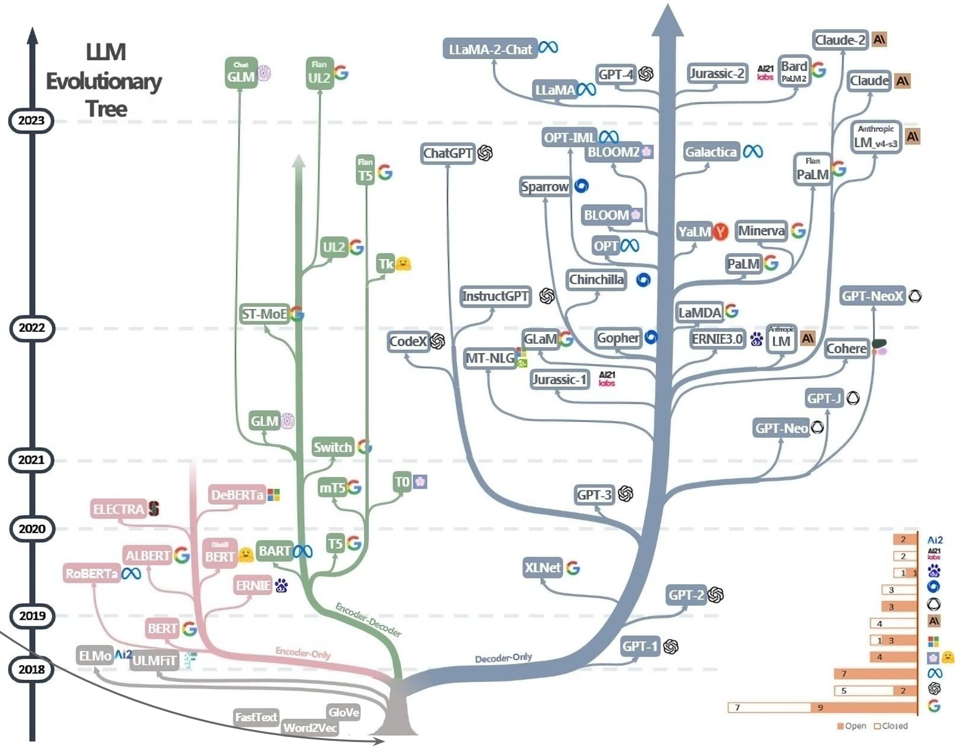
How do we go from purpose driven models to LLMs?

Self-Supervised Learning

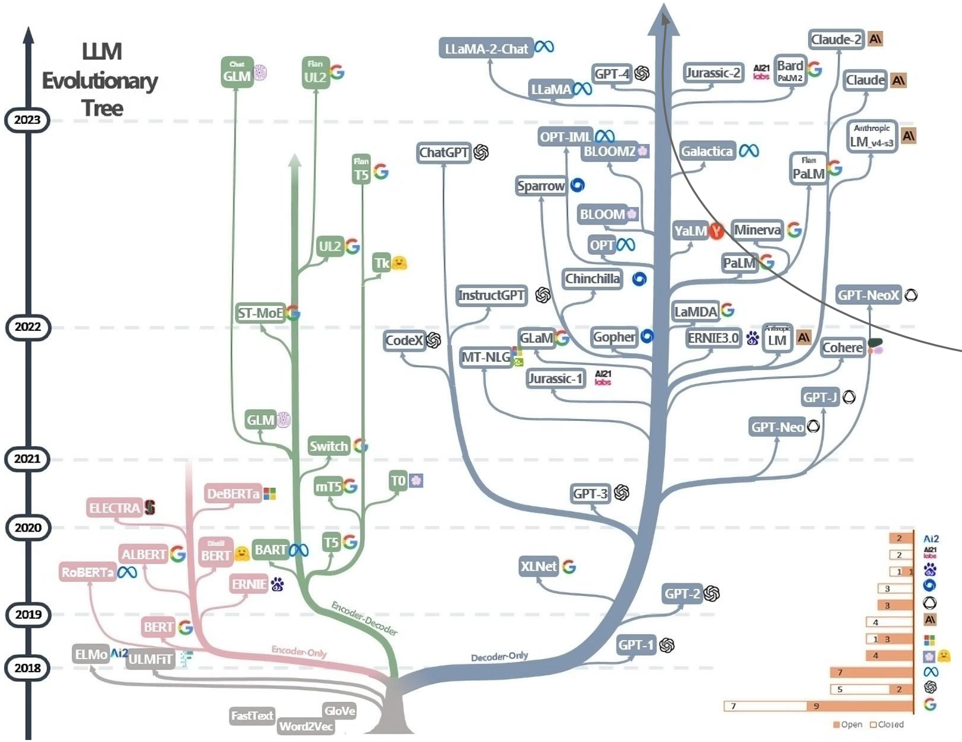
How do we most effectively turn raw text into meaningful loss?

<u>Covered Today (& In Homework)</u>

- Encoder Only
- Decoder Only
- Encoder-Decoder



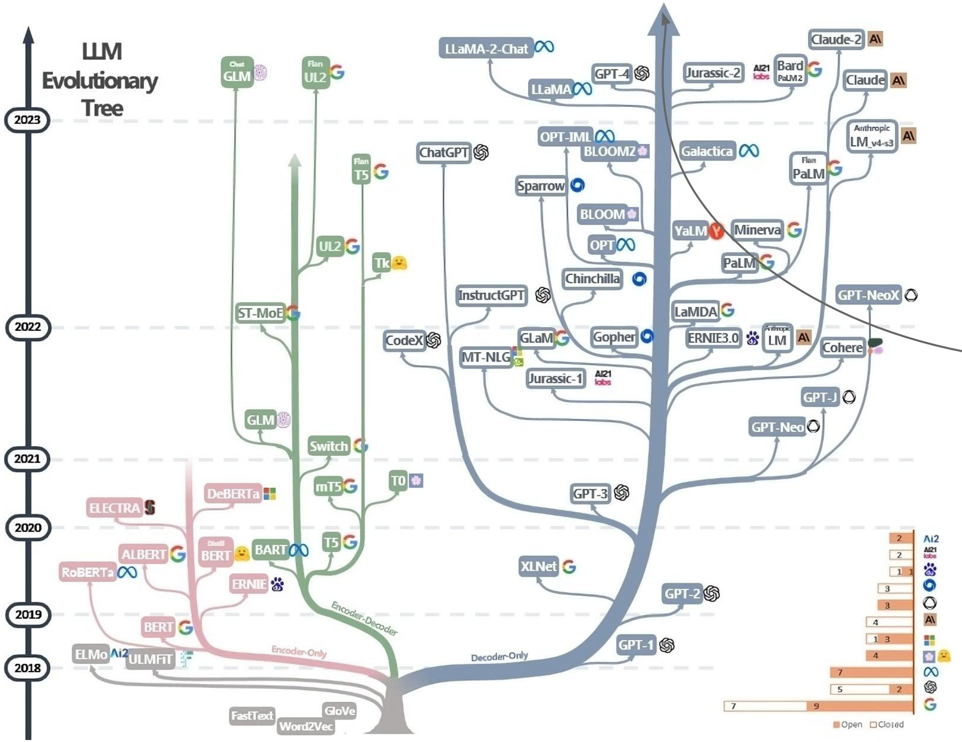
How do we go from purpose driven models to LLMs?



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?

<u>Covered Today</u>

- Data Curation Over Time
- Distributed Training
- “Alignment”

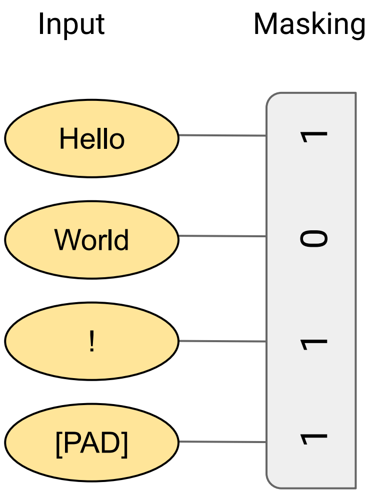
https://github.com/Mooler0410/LLMsPracticalGuide

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?

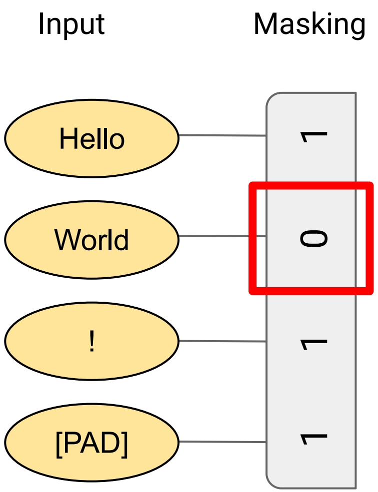
SSL | From raw text to loss!



Masked Language Model

[Devlin et al. 2018 (BERT)]

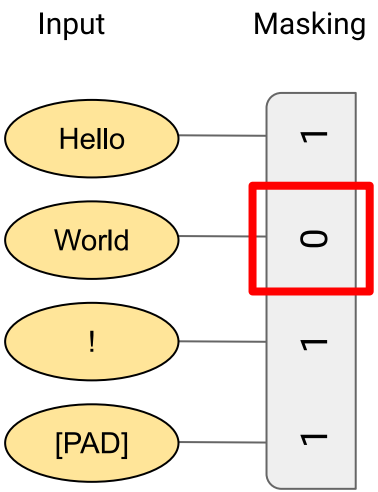
SSL | What is the “Mask” in a Masked Language Model?



Masked Language Model

Devlin et al. 2018 (BERT)

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



Masked Attention

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

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

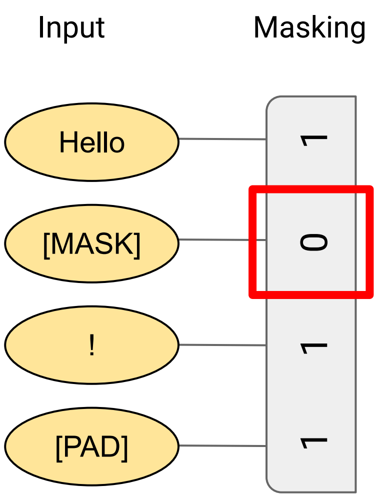
Output vectors: Y = AX

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

Masked Language Model

Devlin et al. 2018 (BERT)

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



Intuition

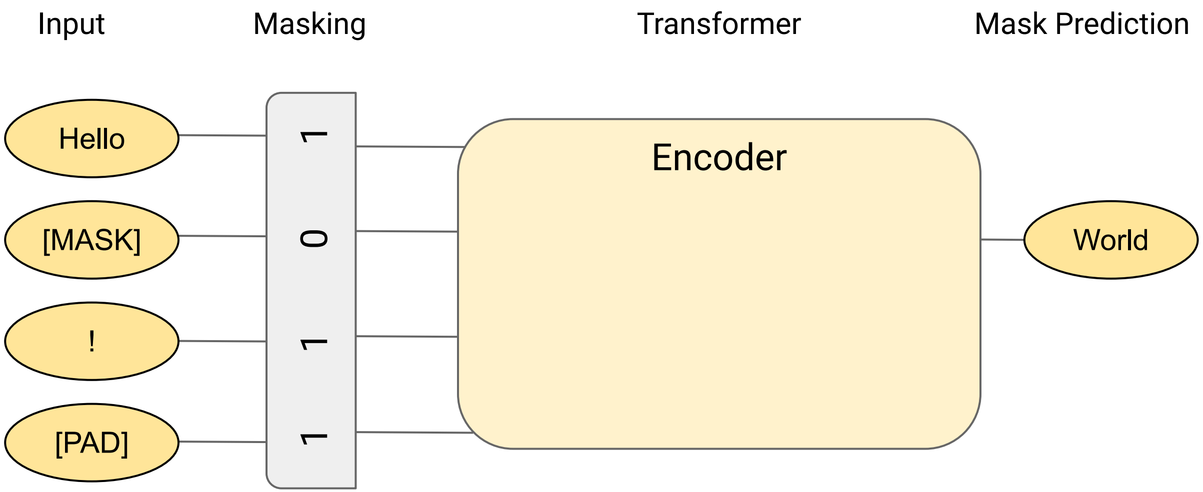
If $MASK_i = 0$, then $Y_i = \sum_{j,j!=i} A_{i,j} X$

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



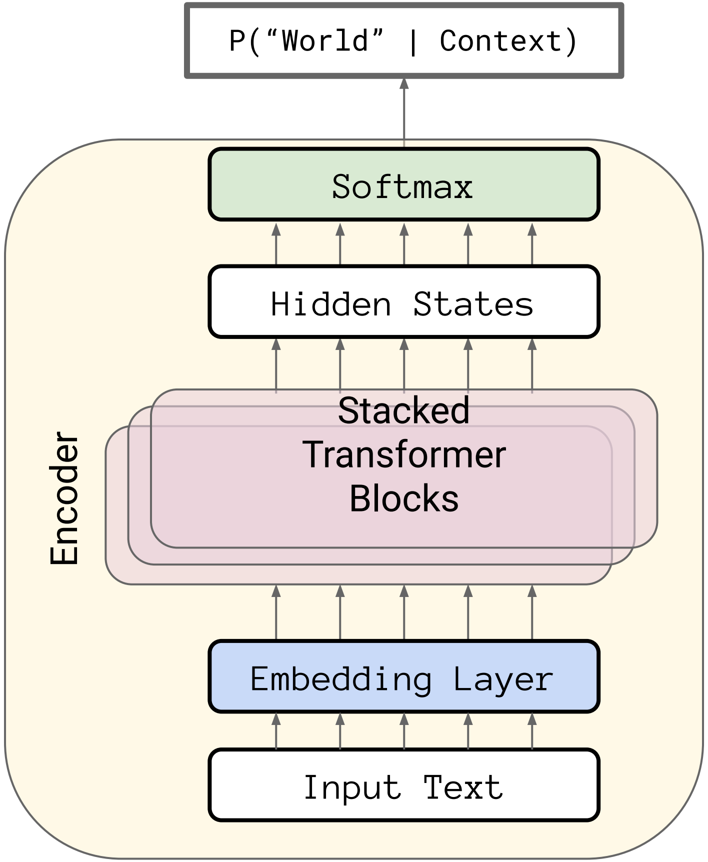
Masked Language Model

[Devlin et al. 2018 (BERT)]

SSL | Masked Token Prediction

Optimize Negative Log Likelihood

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

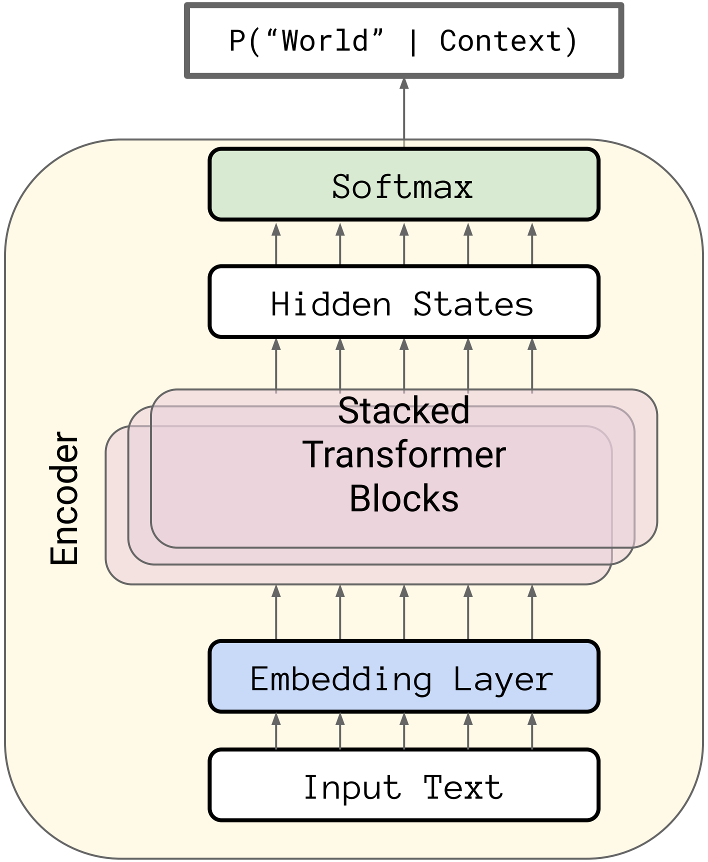


SSL | Masked Token Prediction

Optimize Negative Log Likelihood

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

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



Data | BERT used existing curation!

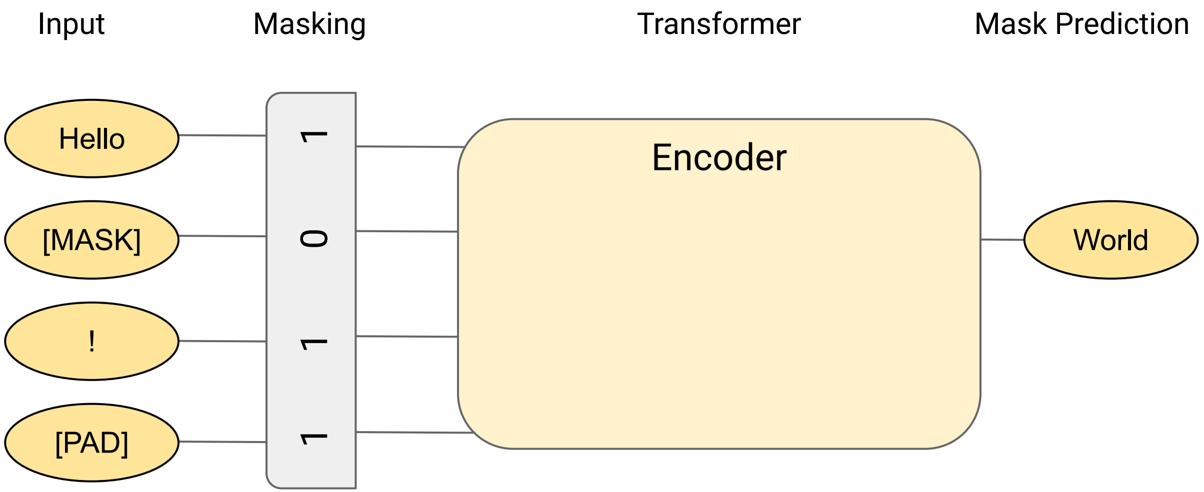
<u>BERT Corpus</u>
English Wikipedia + BooksCorpus

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

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

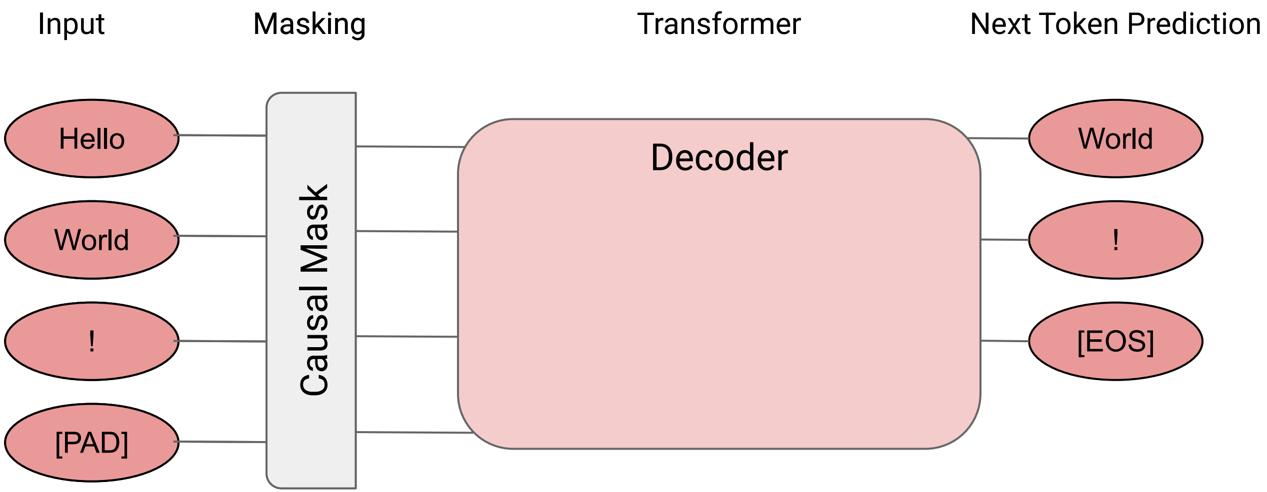
Devlin et al. 2018 (BERT)

Questions?



Masked Language Model

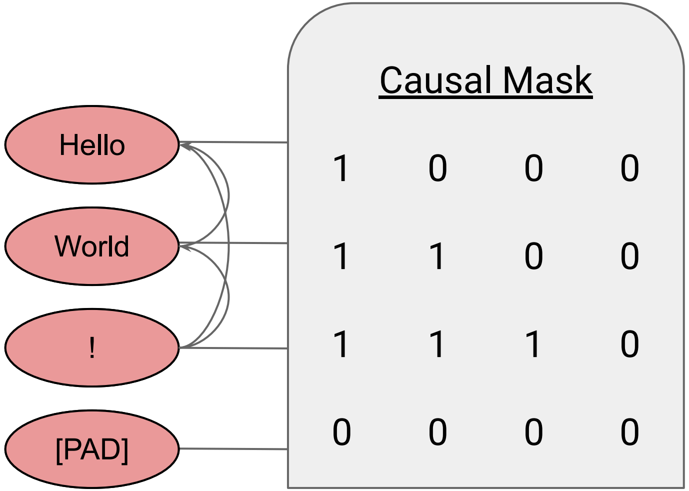
SSL | “How does GPT work?”



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

SSL | Autoregressive Language Modeling

Masking



Masked Attention Again!

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

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

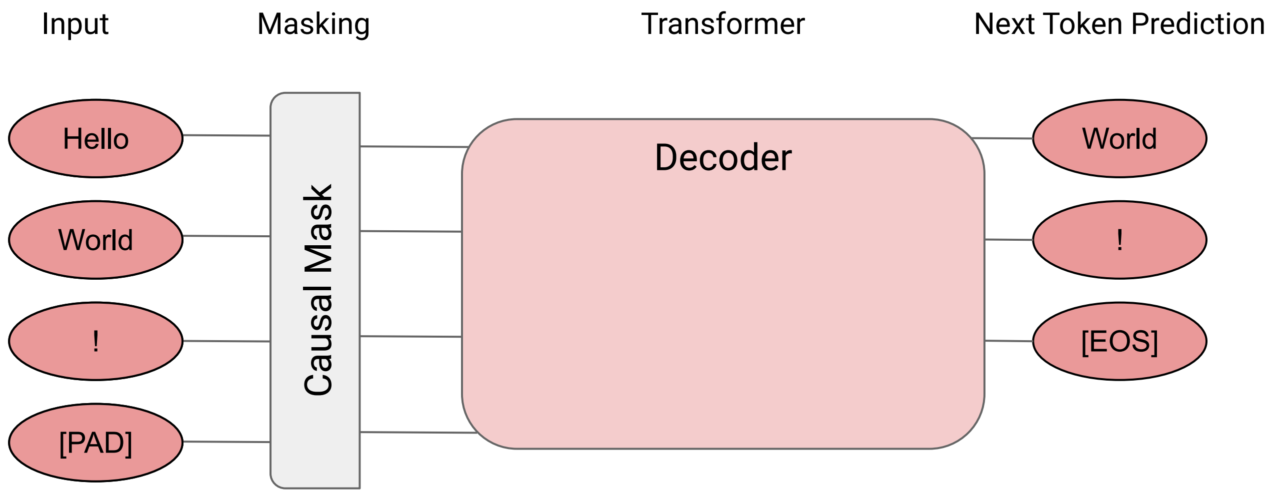
`Output vectors: Y = AX`

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

Tokens only affected by preceding tokens

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

SSL | 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

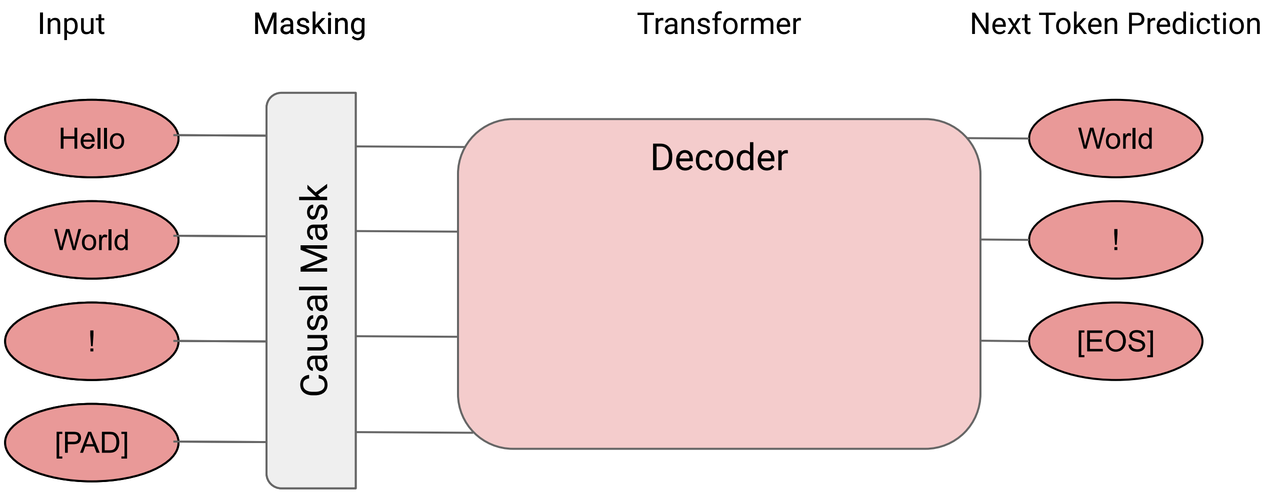
Size
~10 Billion Tokens

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)]

Questions?



Autoregressive Language Model

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

+ Generative Abilities

This is pretty essential!

- Retrieval
- Classification

[Wang et al. 2022]

SSL | Encoder-Only vs. Decoder-Only

Encoder

How to keep this?

+ Retrieval
+ Classification
- No Generative Abilities

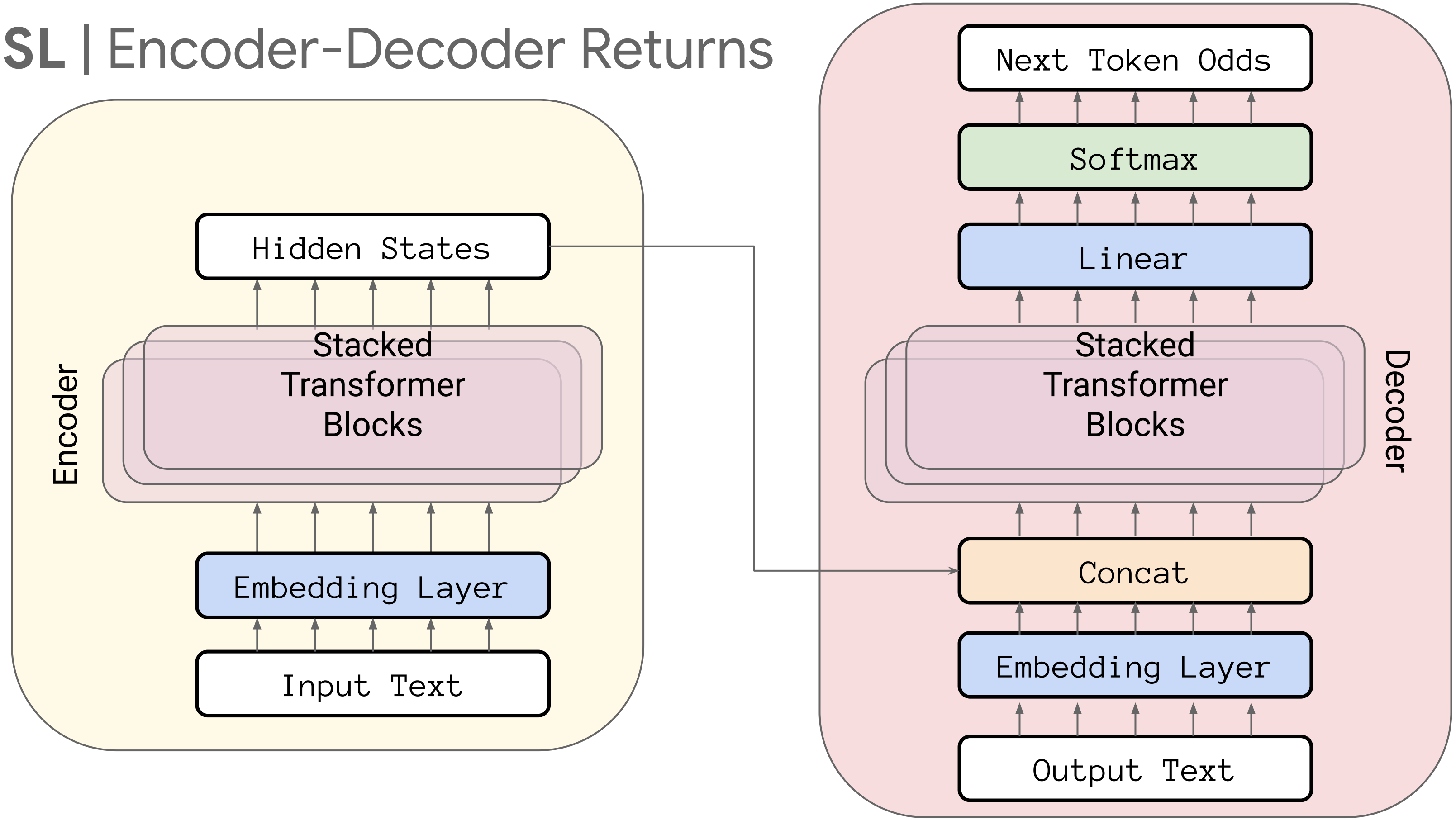
Decoder

+ Generative Abilities

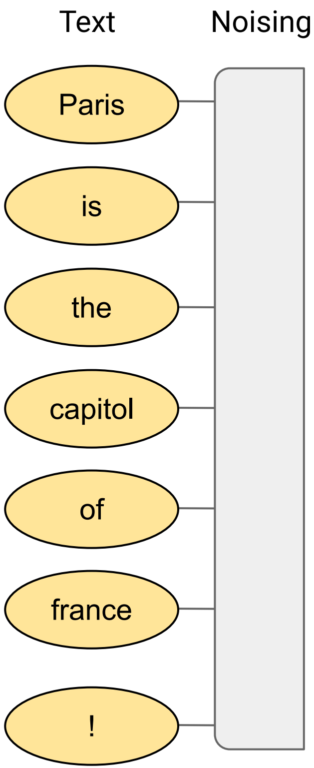
This is pretty essential!

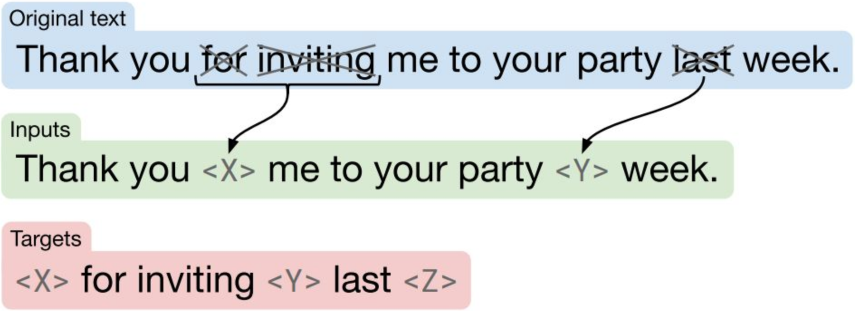
- Retrieval
- Classification

SSL | Encoder-Decoder Returns



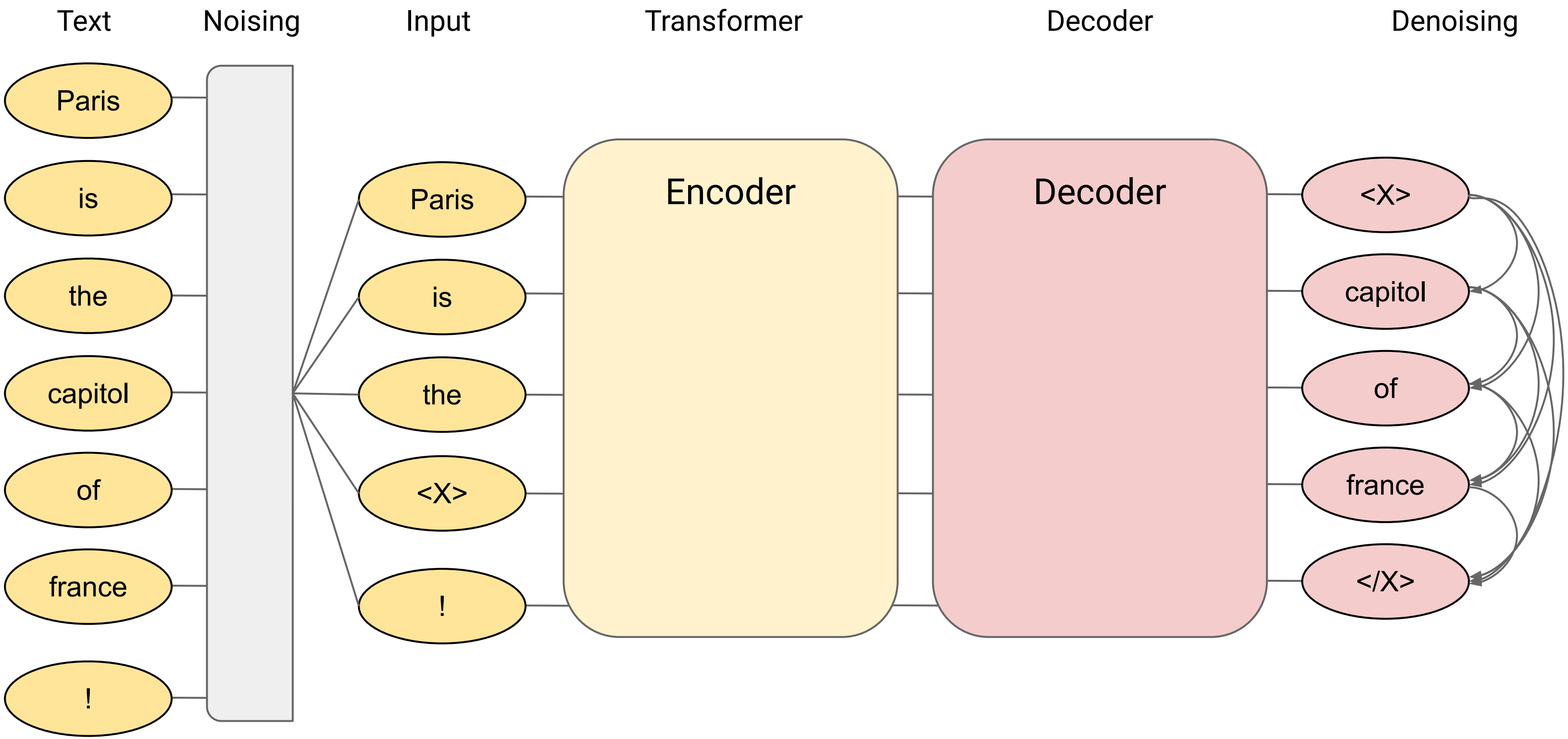
SSL | Universal Text-to-Text





[Raffel et al. 2019]

SSL | Universal Text-to-Text



[Raffel et al. 2019]

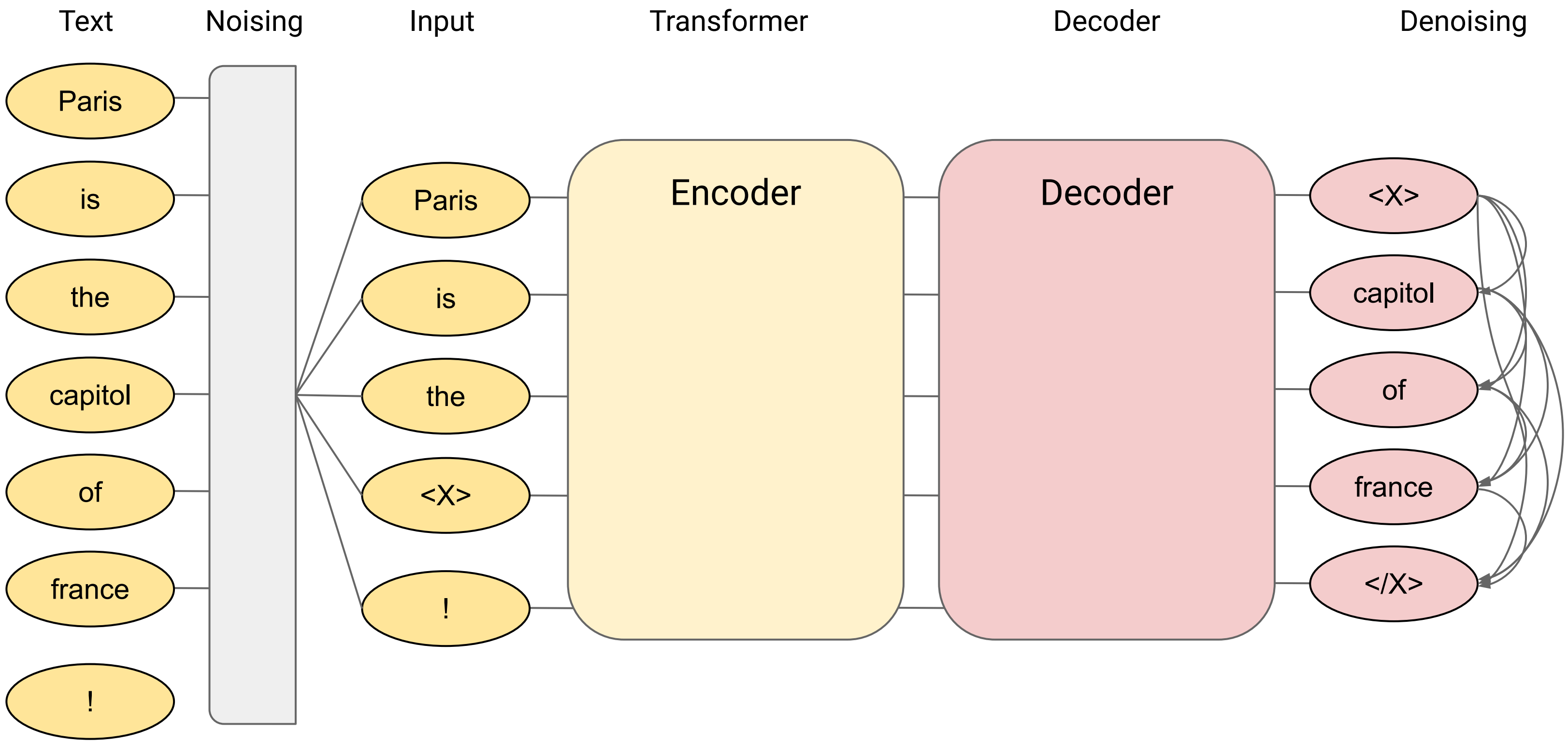
SSL | Universal Text-to-Text

Regardless of noise, Loss Function remains the same still!

Continue using Negative Log Likelihood

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

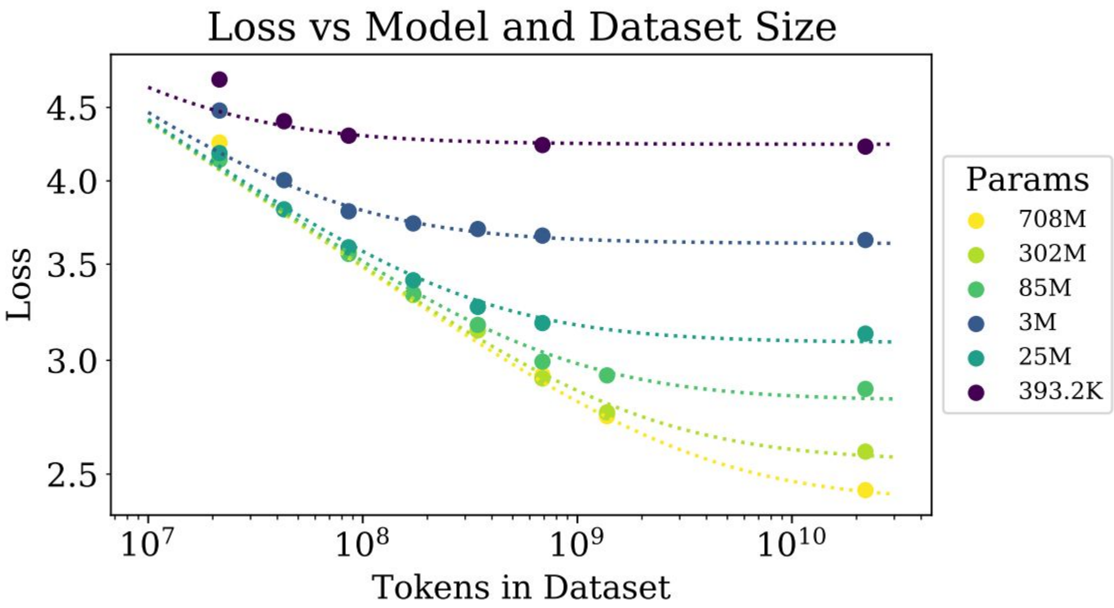
Questions?



Text
Noising
Input
Transformer
Decoder
Denoising
Paris
is
the
capitol
of
france
!
Paris
is
the
<X>
!
Encoder
Decoder
<X>
capitol
of
france
</X>

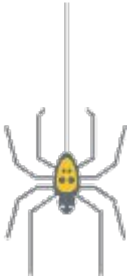
Data & Parameter Scaling | Moving to **Large** Language Models

Today's LLMs are driven data and model scaling



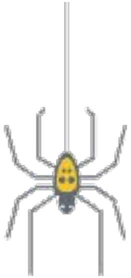
[Kaplan et al. 2020]

Data Scaling | Collecting High-Quality Self-Supervision at Scale



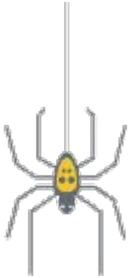
We could get a lot more data from CommonCrawl!

Data Scaling | Collecting High-Quality Self-Supervision at Scale



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

Data Scaling | Collecting High-Quality Self-Supervision at Scale



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

T5 - Encoder-Decoder with Common Crawl Scale Data

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

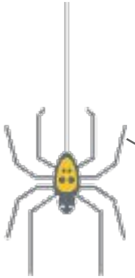
Size
~350 Billion Tokens

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.

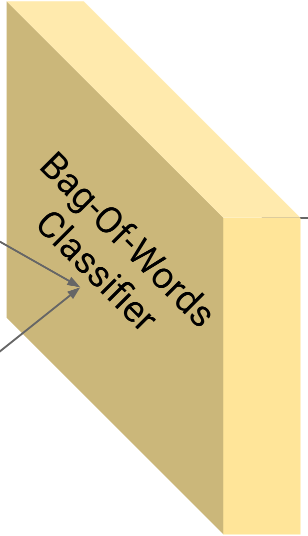
Raffel et al. 2019

GPT-3 - Increased Scaling Via Automated Data Curation



Low-Quality, High Volume

Training



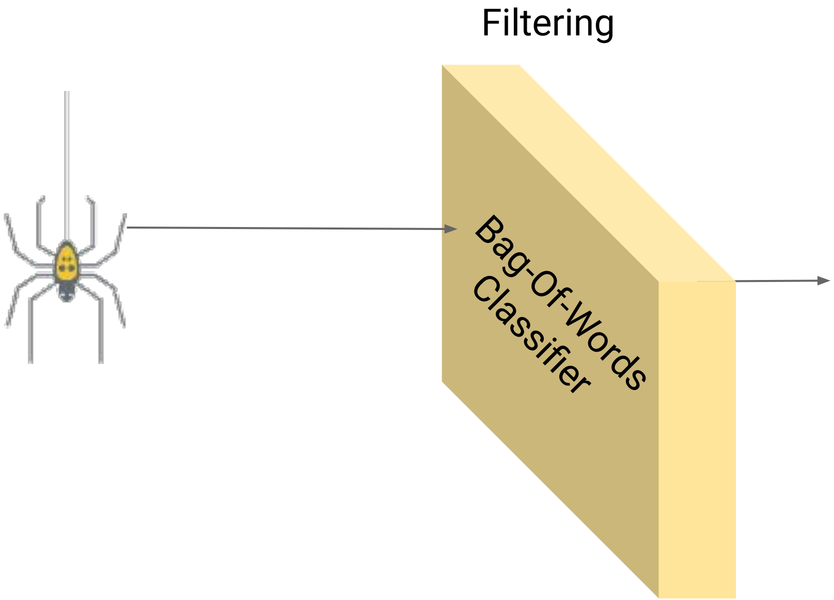
Distinguish High and Low Quality

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

Brown et al. 2020

GPT-3 - Increased Scaling Via Automated Data Curation



Keep “False” Positives

“False” positive ~= High Quality

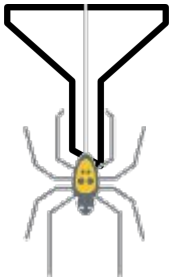
Brown et al. 2020

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 | Data Mixture has become the biggest “secret”

Llama 2 Corpus

Size
> 2 Trillion Tokens

Quality
Minimal details known

Touvron et al. 2023 (b)

Gemini Corpus

Size
Unknown

Quality
No details known

Anil et al. 20232

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)

Gemini Corpus

Size
Unknown

Quality
No details known

Anil et al. 202320

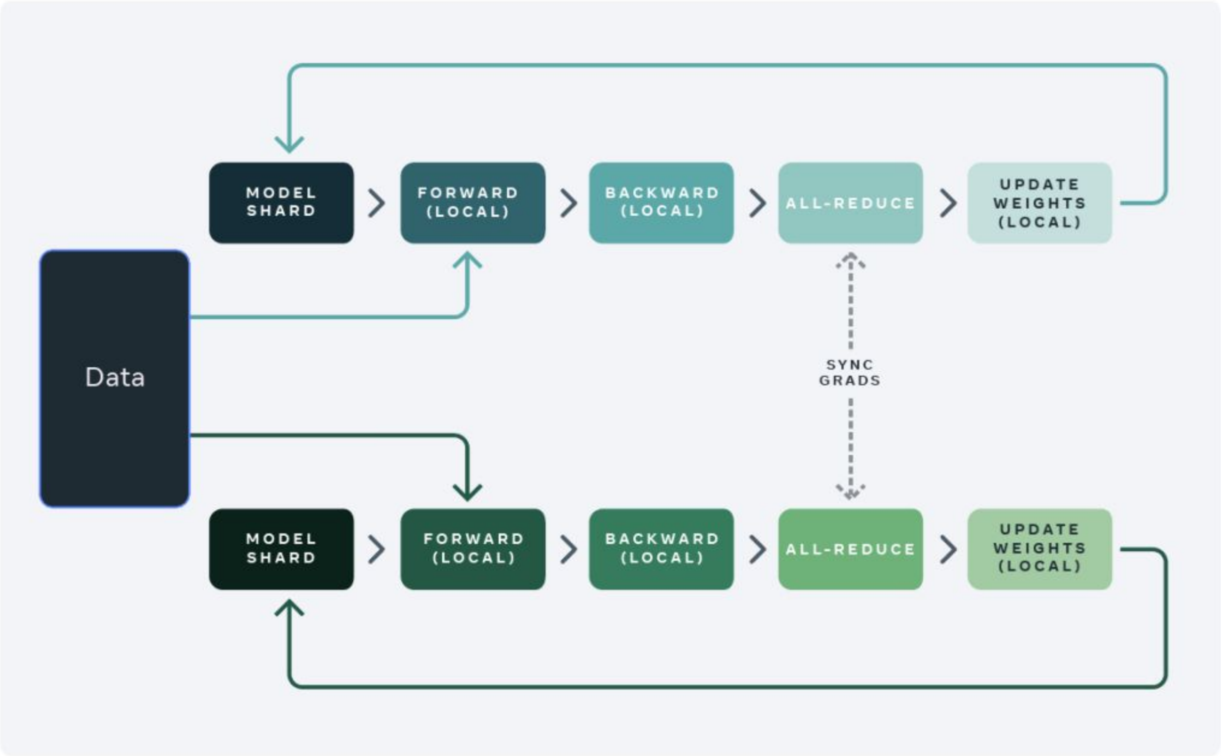
GPT-4 Corpus

Size
Unknown (Est. 11T Tokens)

Quality
No details known

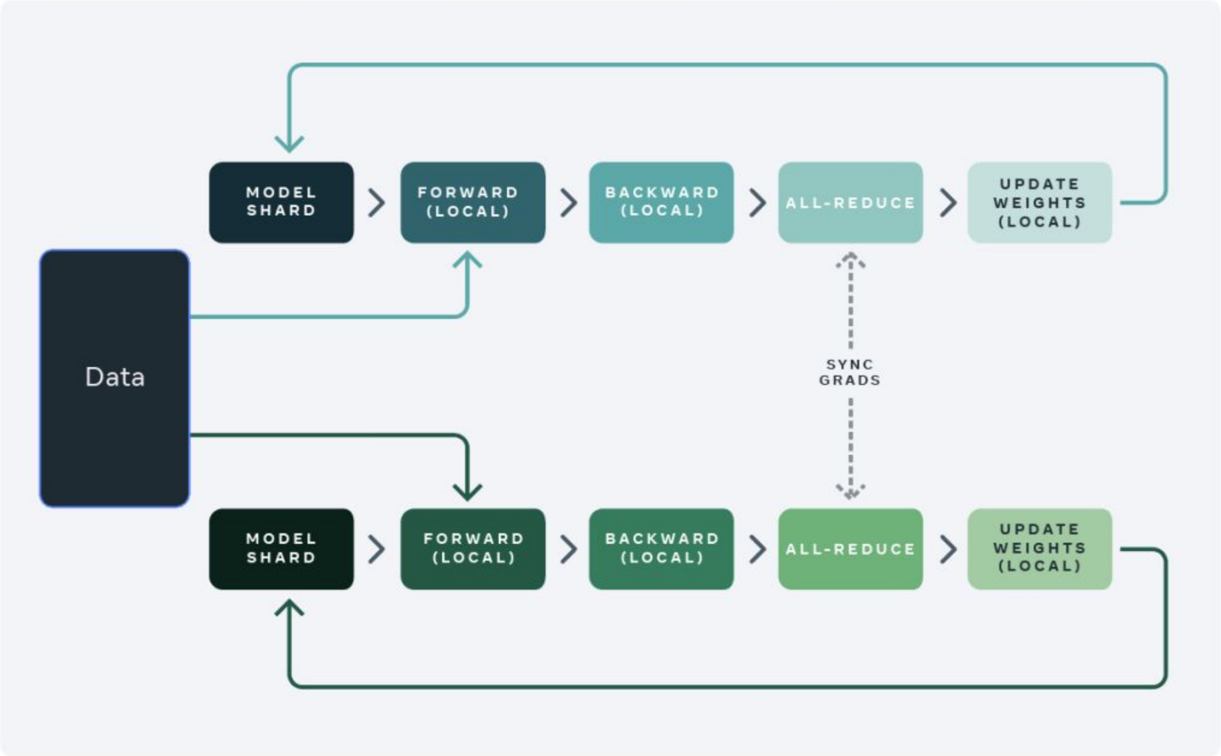
OpenAI 2023

Scaling Parameters | Data Parallel Training



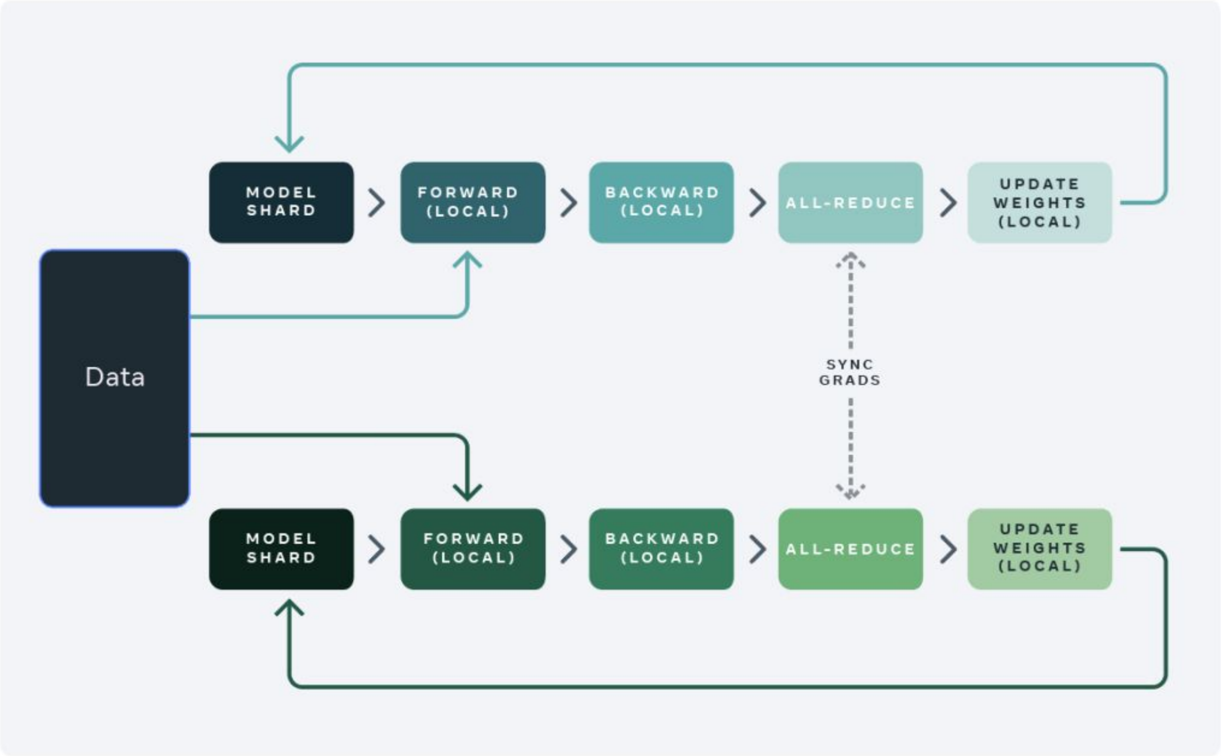
https://engineering.fb.com/2021/07/15/open-source/fsdp/

Scaling Parameters | Data Parallel Training



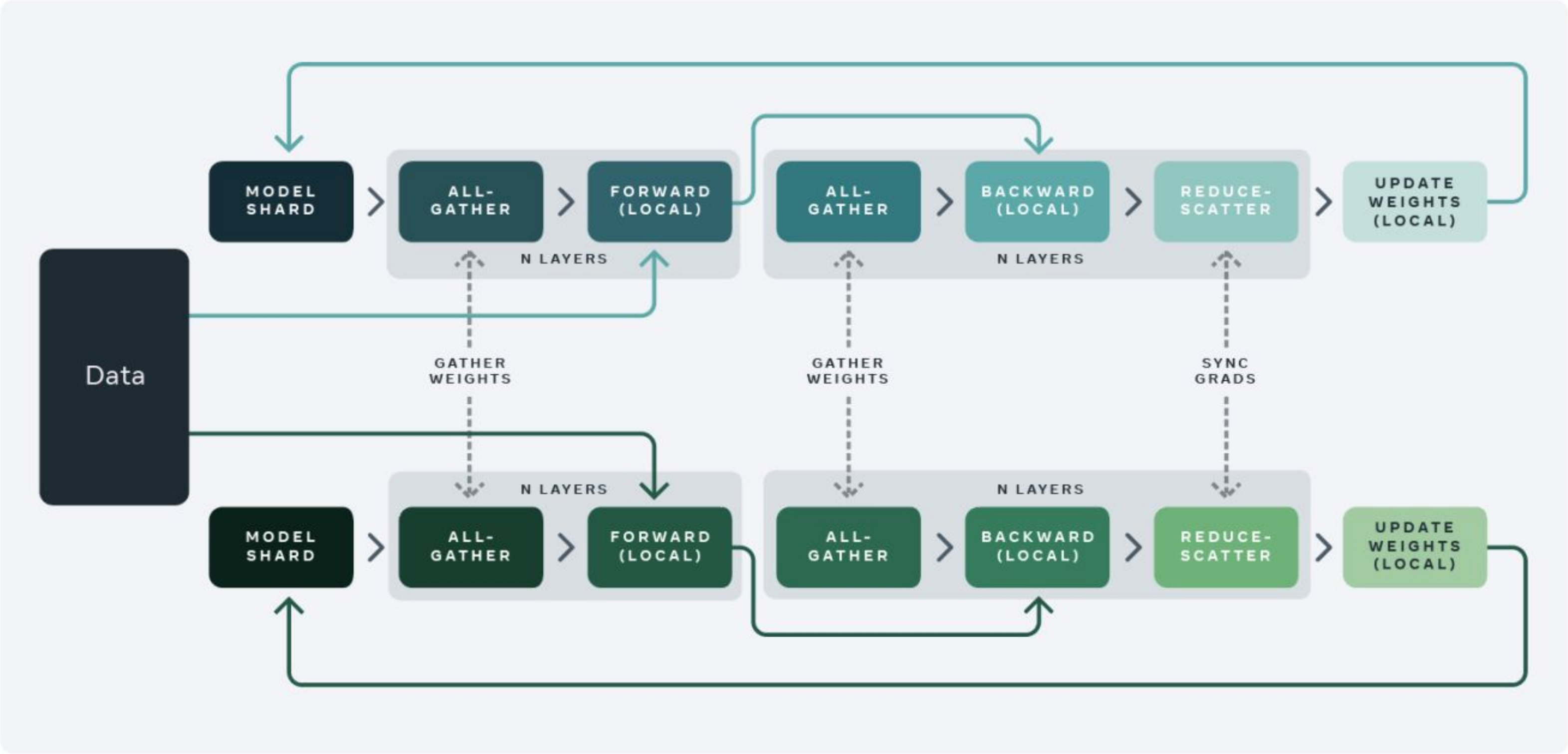
Total memory increases linearly with shards

Scaling Parameters | Data Parallel Training



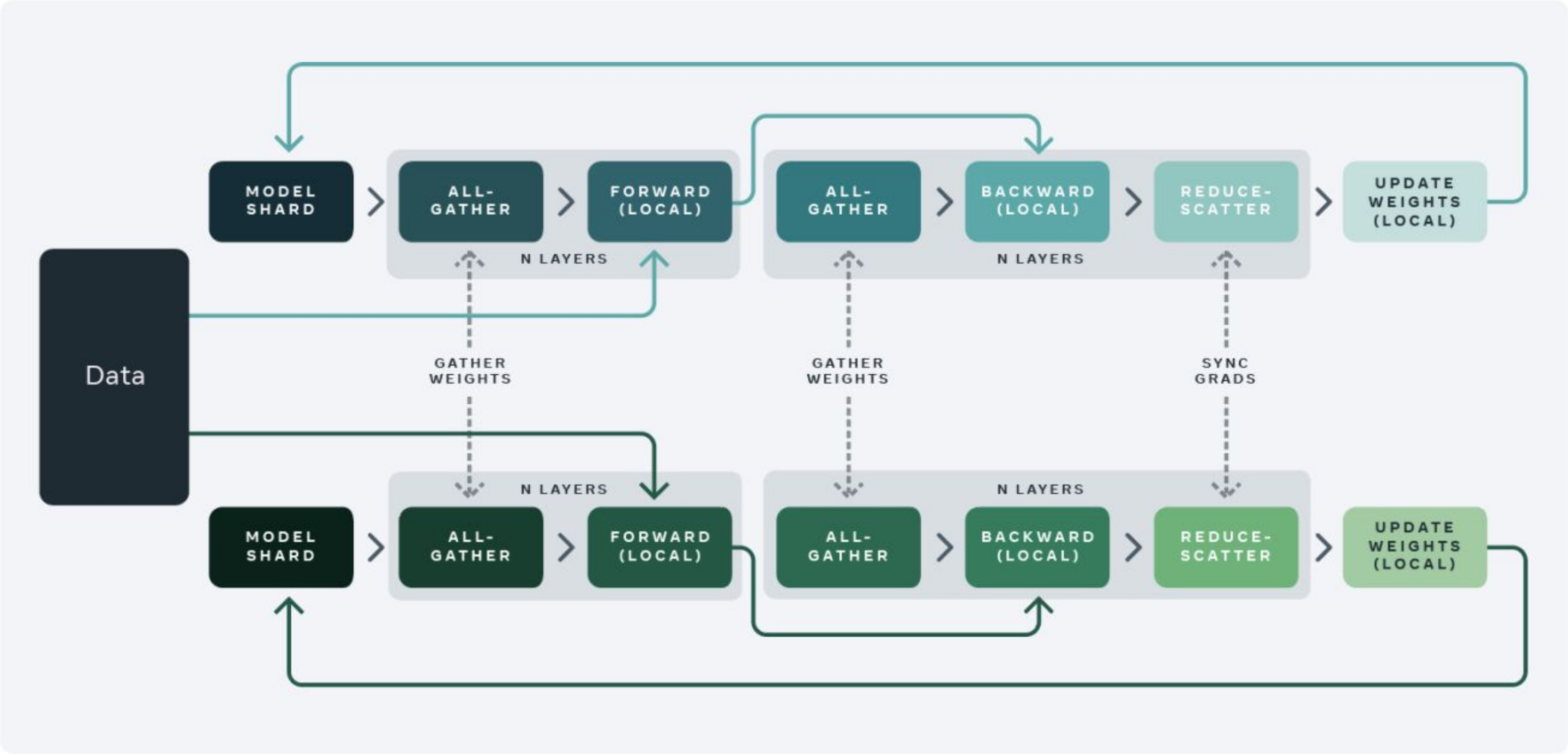
Max memory constrains model size

Scaling Parameters | *Fully* Sharded Data Parallel Training



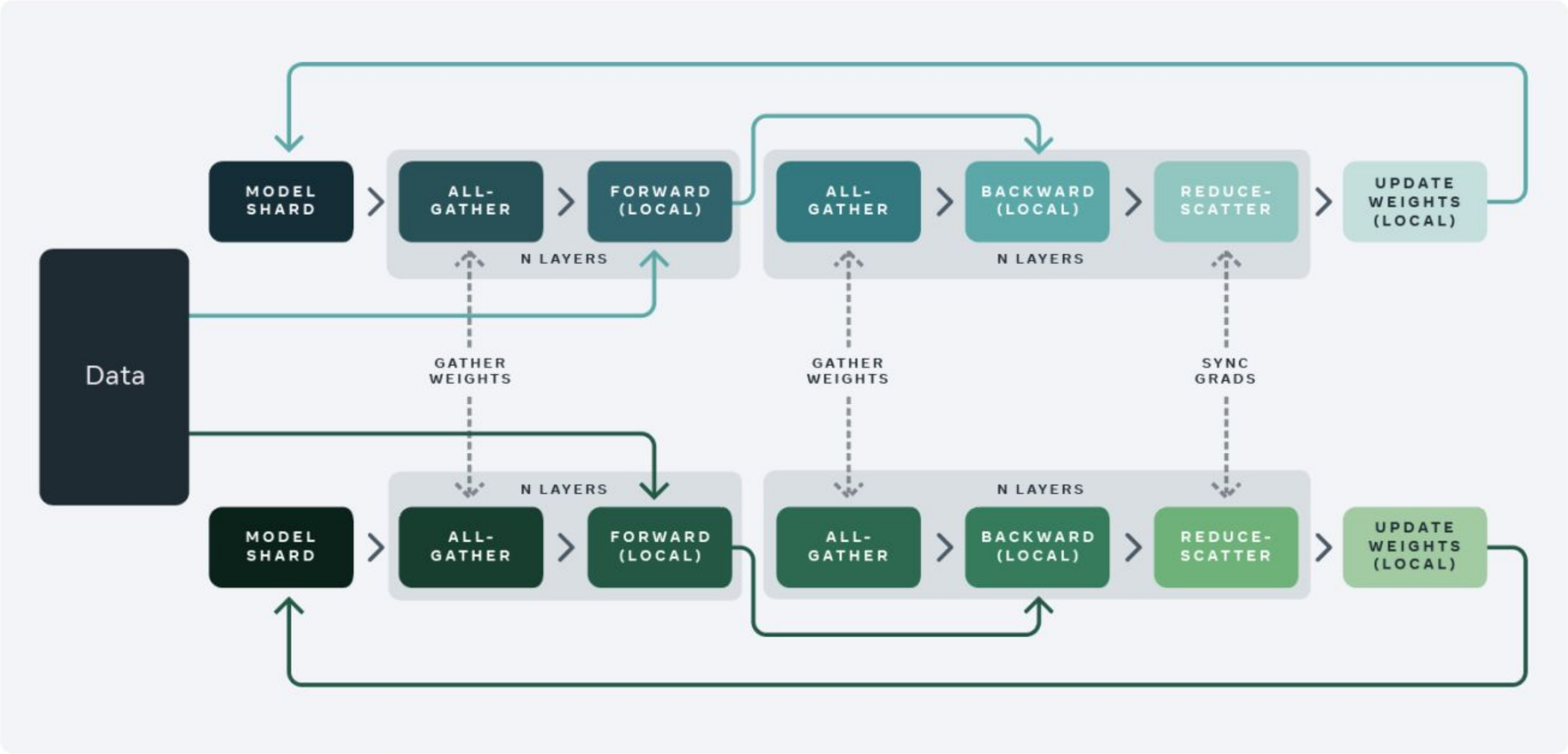
https://engineering.fb.com/2021/07/15/open-source/fsdp/

Scaling Parameters | *Fully* Sharded Data Parallel Training



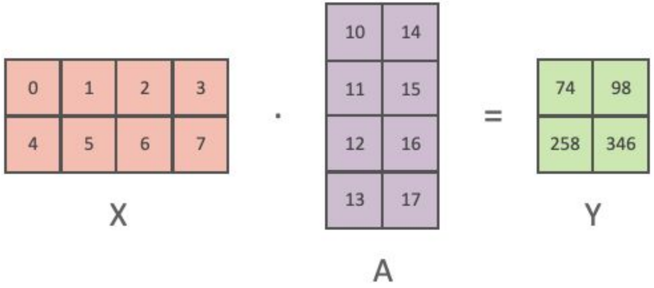
Total memory is constant

Scaling Parameters | *Fully* Sharded Data Parallel Training



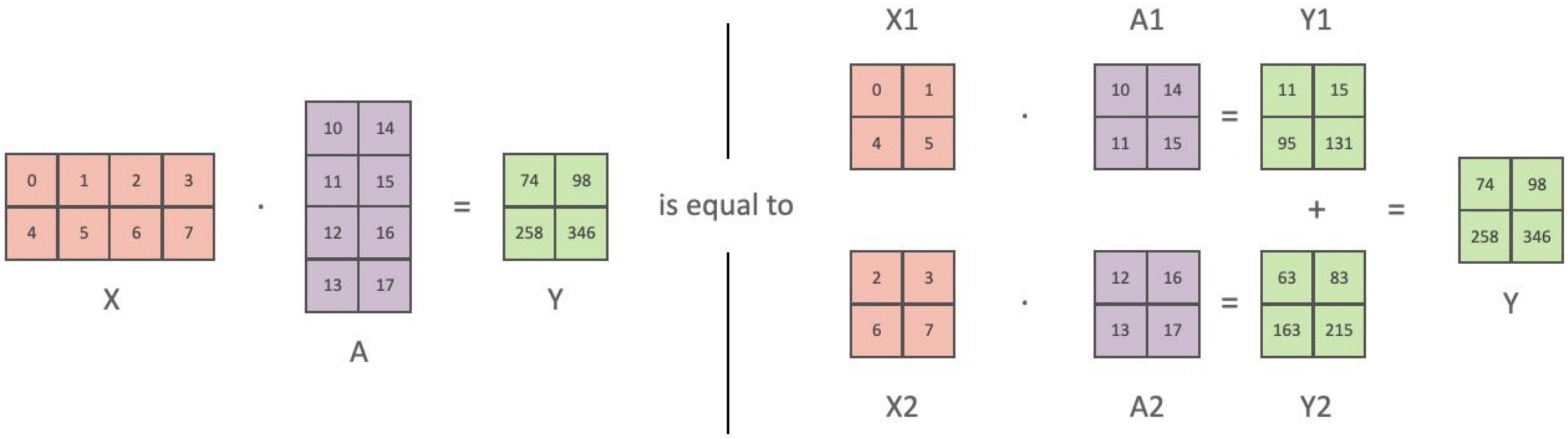
Max single GPU memory constrains layer size

Scaling Parameters | Tensor Parallel Training



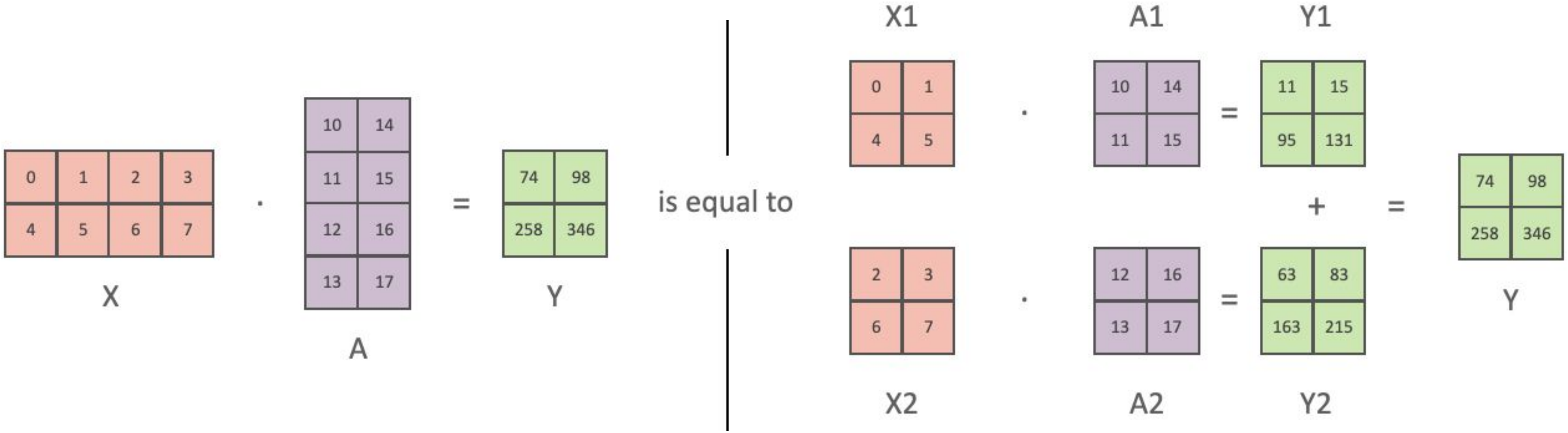
https://huggingface.co/docs/transformers/v4.15.0/parallelism#tensor-parallelism

Scaling Parameters | Tensor Parallel Training



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

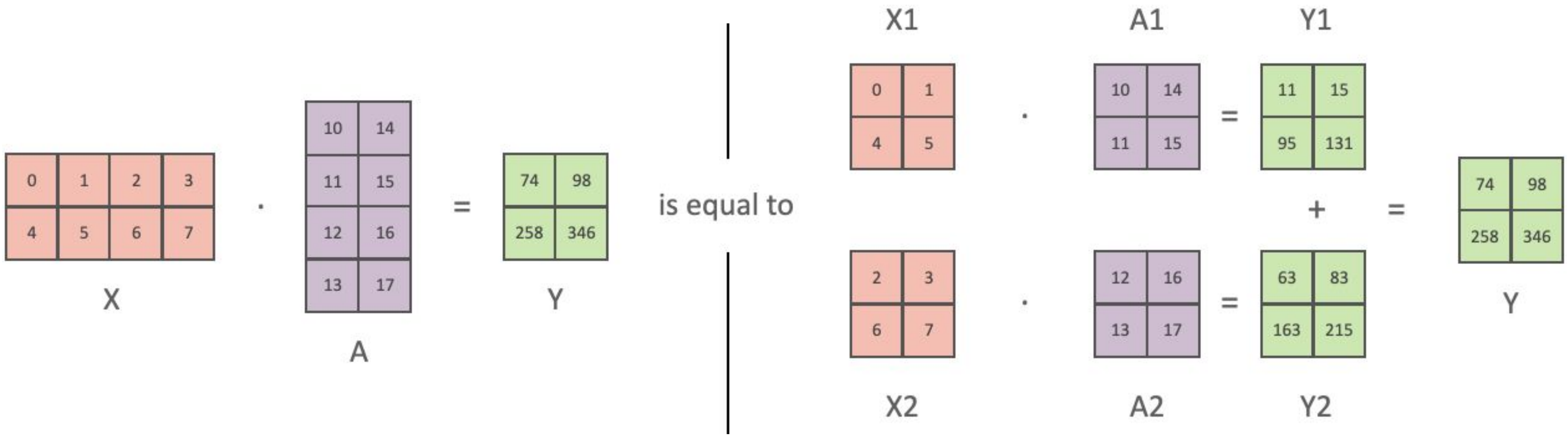
Scaling Parameters | Tensor Parallel Training



Don't need to sync gradients!

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

Scaling Parameters | Tensor Parallel Training

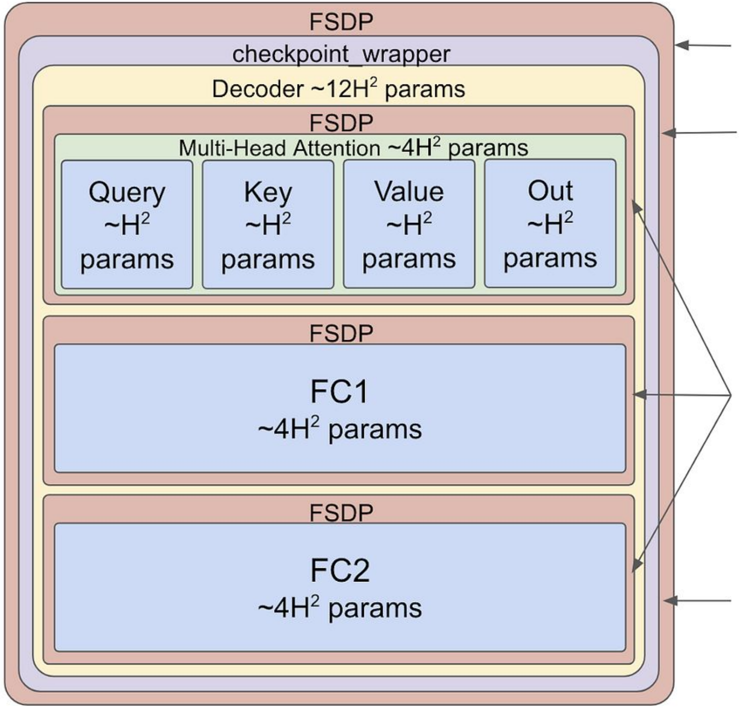


Don't need to sync gradients!

Max GPU memory constrains layer shard size

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

Scaling Parameters | FSDP + TP = ~Limitless Scaling



Outer activations are offloaded to CPU

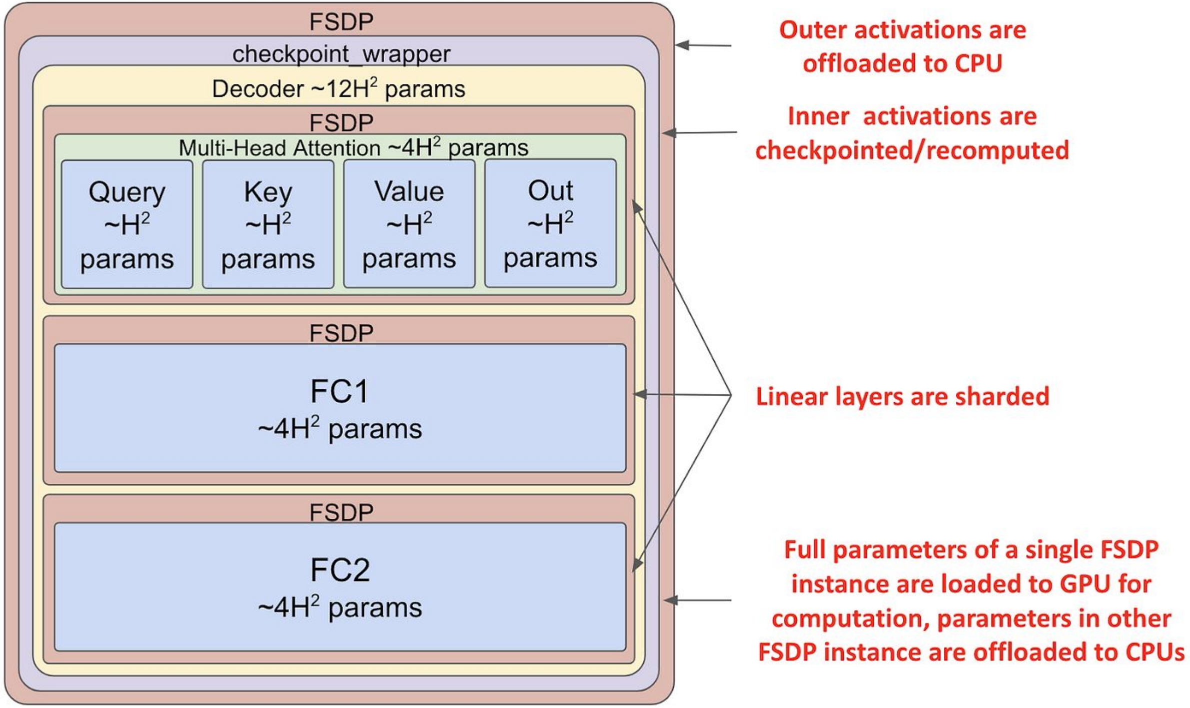
Inner activations are checkpointed/recomputed

Linear layers are sharded

Full parameters of a single FSDP instance are loaded to GPU for computation, parameters in other FSDP instance are offloaded to CPUs

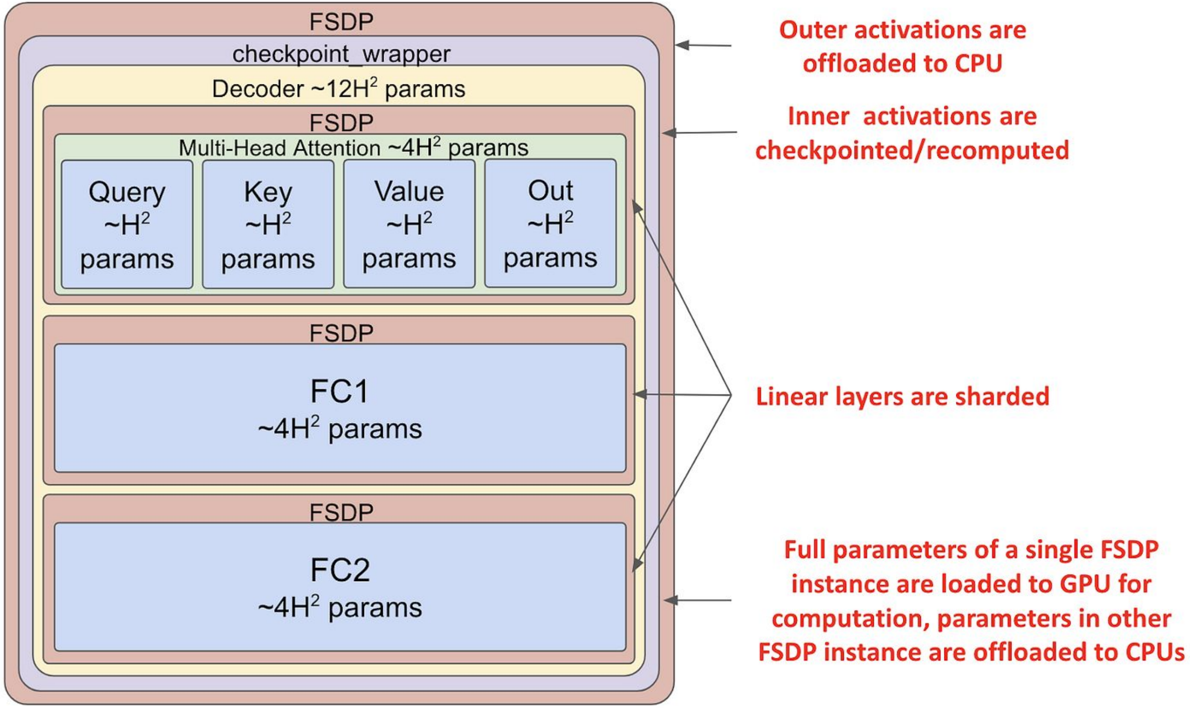
[1 Trillion Parameter Model with Tensor Parallelism and FSDP]

Scaling Parameters | FSDP + TP = ~Limitless Scaling



[1 Trillion Parameter Model with Tensor Parallelism and FSDP]

Questions?


FSDP
checkpoint_wrapper
Decoder ~12H2 params
FSDP
Multi-Head Attention ~4H2 params
Query ~H2 params
Key ~H2 params
Value ~H2 params
Out ~H2 params
FSDP
FC1 ~4H2 params
FSDP
FC2 ~4H2 params
Outer activations are offloaded to CPU
Inner activations are checkpointed/recomputed
Linear layers are sharded
Full parameters of a single FSDP instance are loaded to GPU for computation, parameters in other FSDP instance are offloaded to CPUs

LLMs aren't very good for "assistance" by default

Suggest 3 things to do in San Mateo.

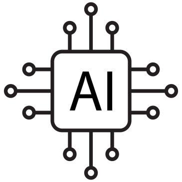
San Mateo is a city in San Mateo County, California, in the high-tech enclave of Silicon Valley in the San Francisco Bay Area. With a population of 97,207 at the 2010 census, the city is located on the San Francisco Peninsula halfway between San Francisco and San Jose, about 20 miles (32 km) from the Pacific Ocean.

Massively Multitask Supervised “Instruction” Finetuning

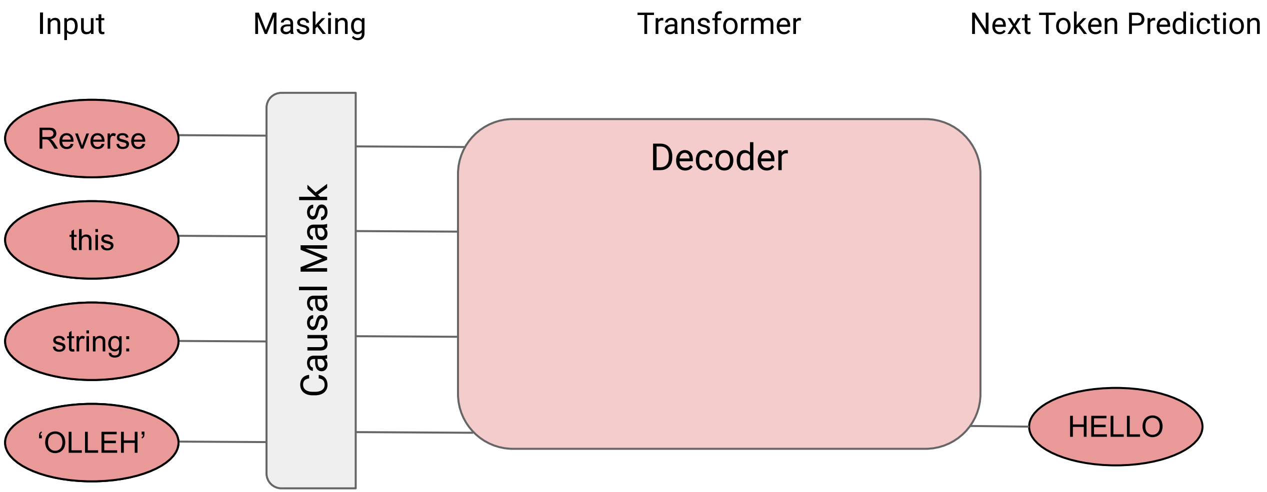
“Reverse this string: ‘OLLEH’.” + “‘HELLO’”

“Suggest 3 things to do in San Mateo.” + “1. Visit San Mateo, Central Park…”

“What is the capital of Qatar?” + “Doha”



Instruction Tuning | Just keep training!

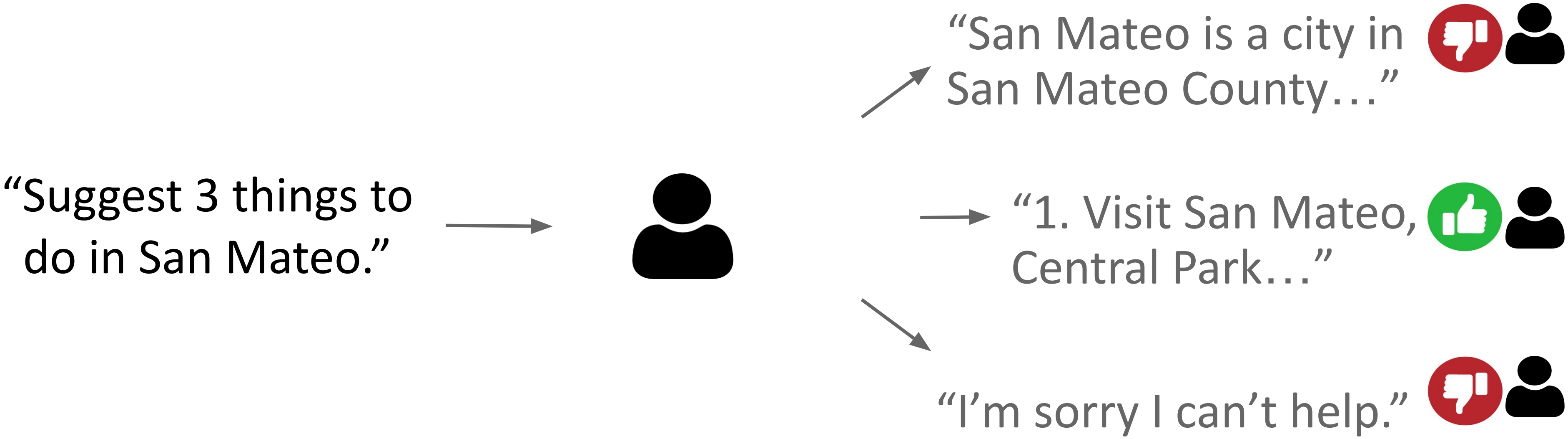


Optimize Negative Log Likelihood of The Response

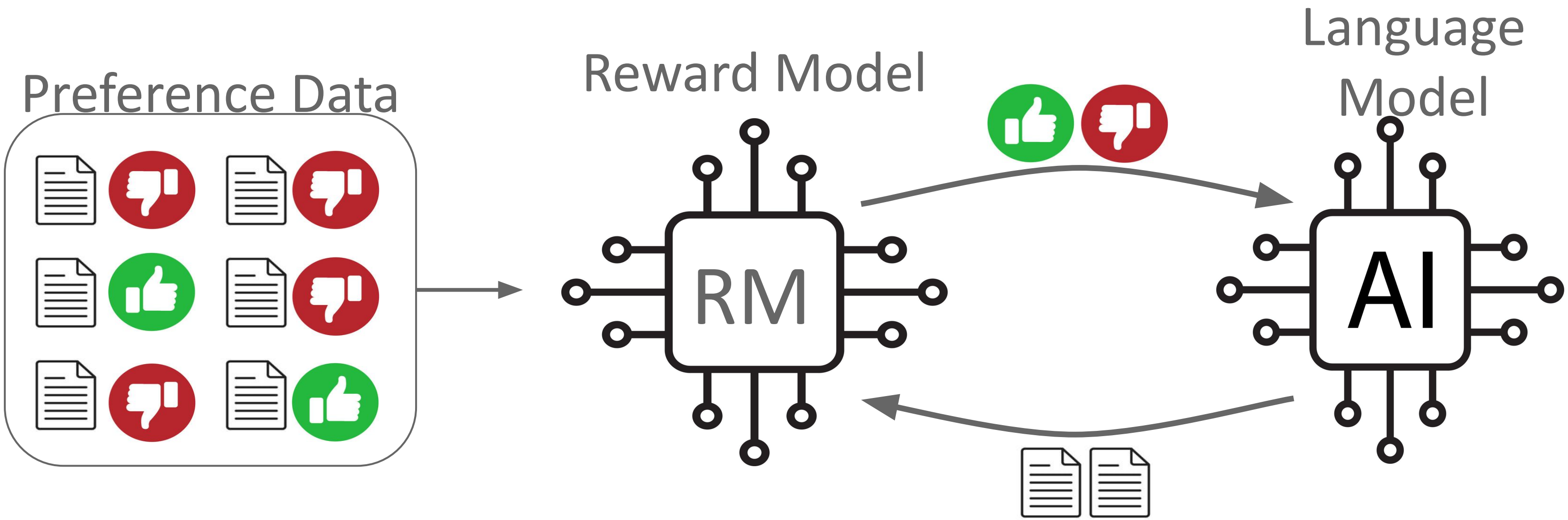
```
loss = -log(P(RESPONSE | INSTRUCTION))
```

[Wang et al. 2022]

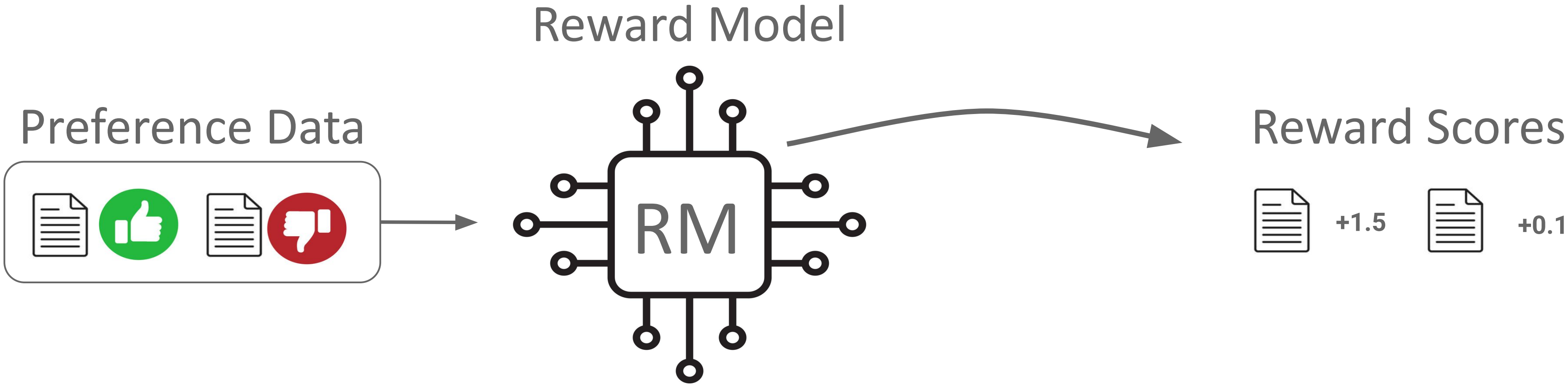
Further Refinement from Sparse Reward (RLHF)



Reinforcement Learning From Human Feedback (RLHF)



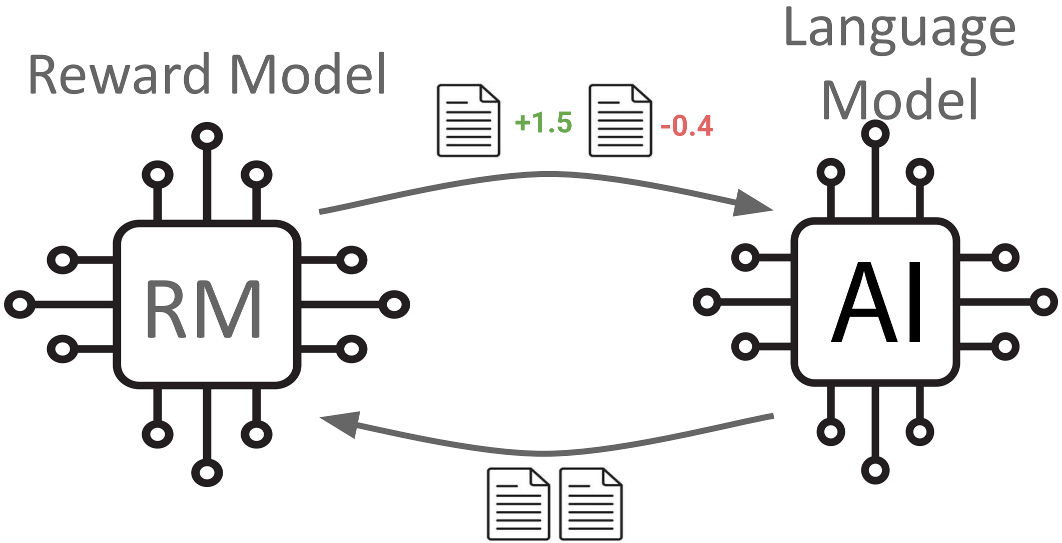
Reinforcement Learning From Human Feedback (RLHF)



Optimize Reward Margin between Preferences

```
loss = -log(σ(RM(POSITIVE) - RM(NEGATIVE)))
```

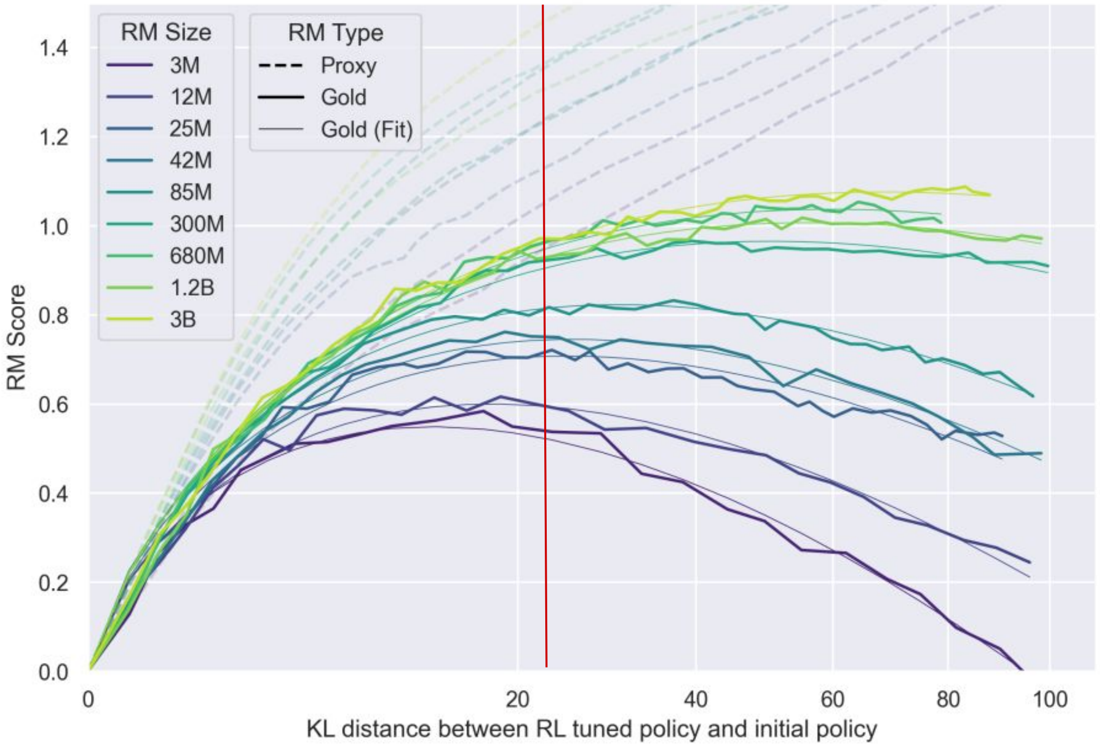
Reinforcement Learning From Human Feedback (RLHF)



Optimize Reward Margin between Preferences

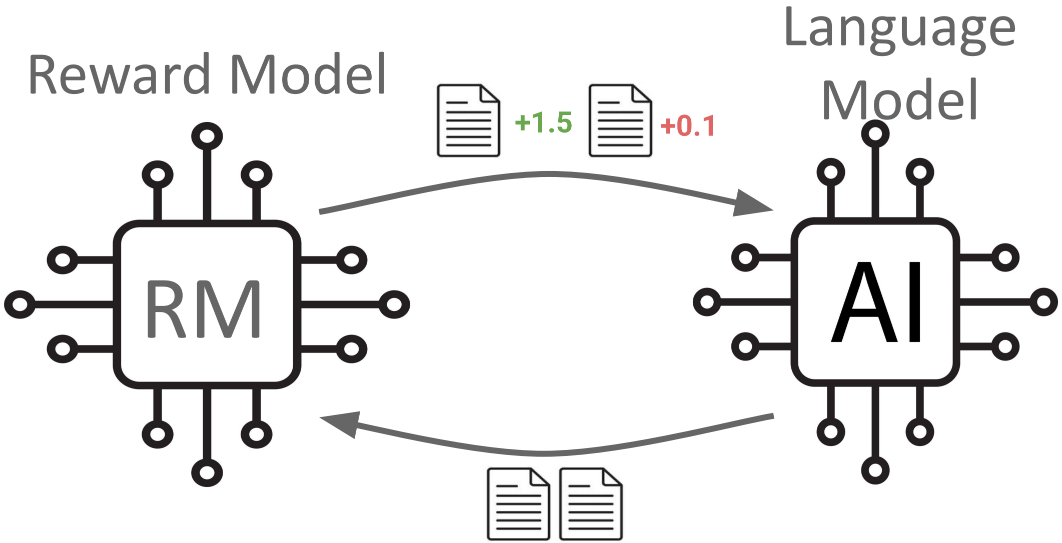
$\texttt{loss}_{\texttt{RM}} = \texttt{-RM(GENERATED\_EXAMPLES)}$

Models Quickly Overfit to Naively Optimized Reward



[Gao et al. 2022]()

Reinforcement Learning From Human Feedback (RLHF)



Optimize Reward Without Drifting Too Far from SFT

$$\text{loss}_{RLHF} = \text{loss}_{RM} + KL(LM_{RLHF}, LM_{SFT})$$

Final Questions?

Fill out my anonymous feedback form