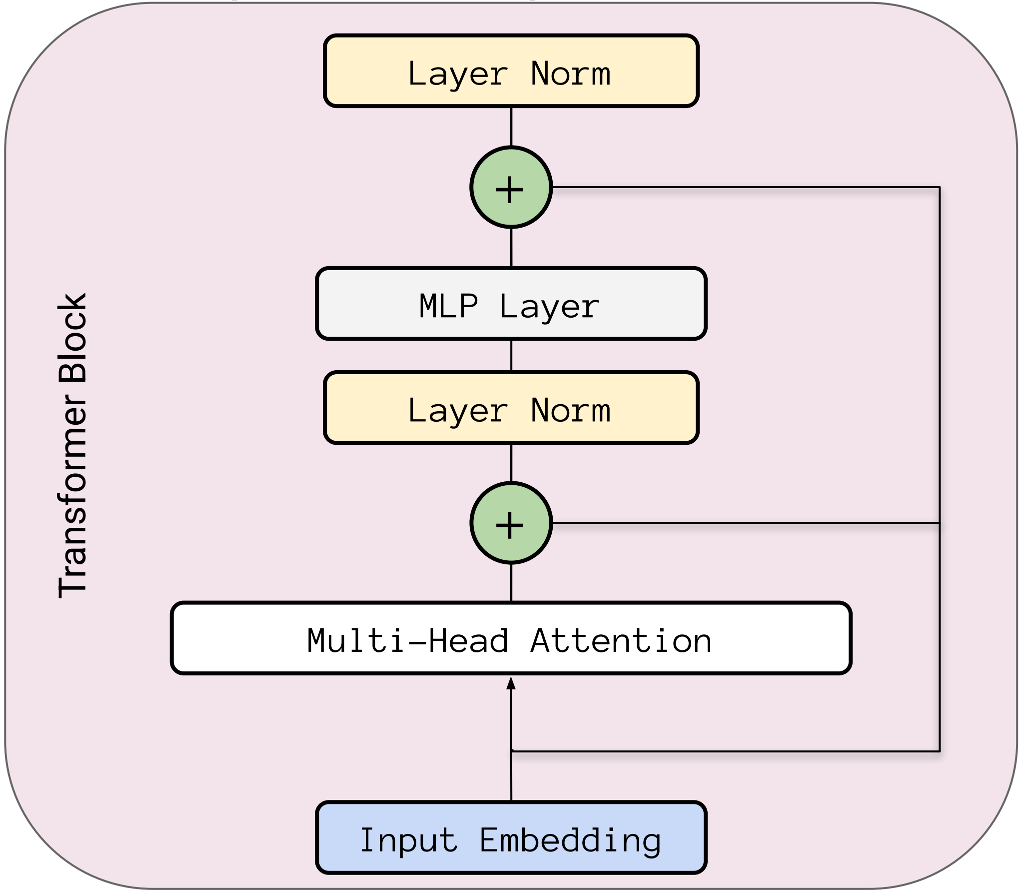
Training Large Language Models

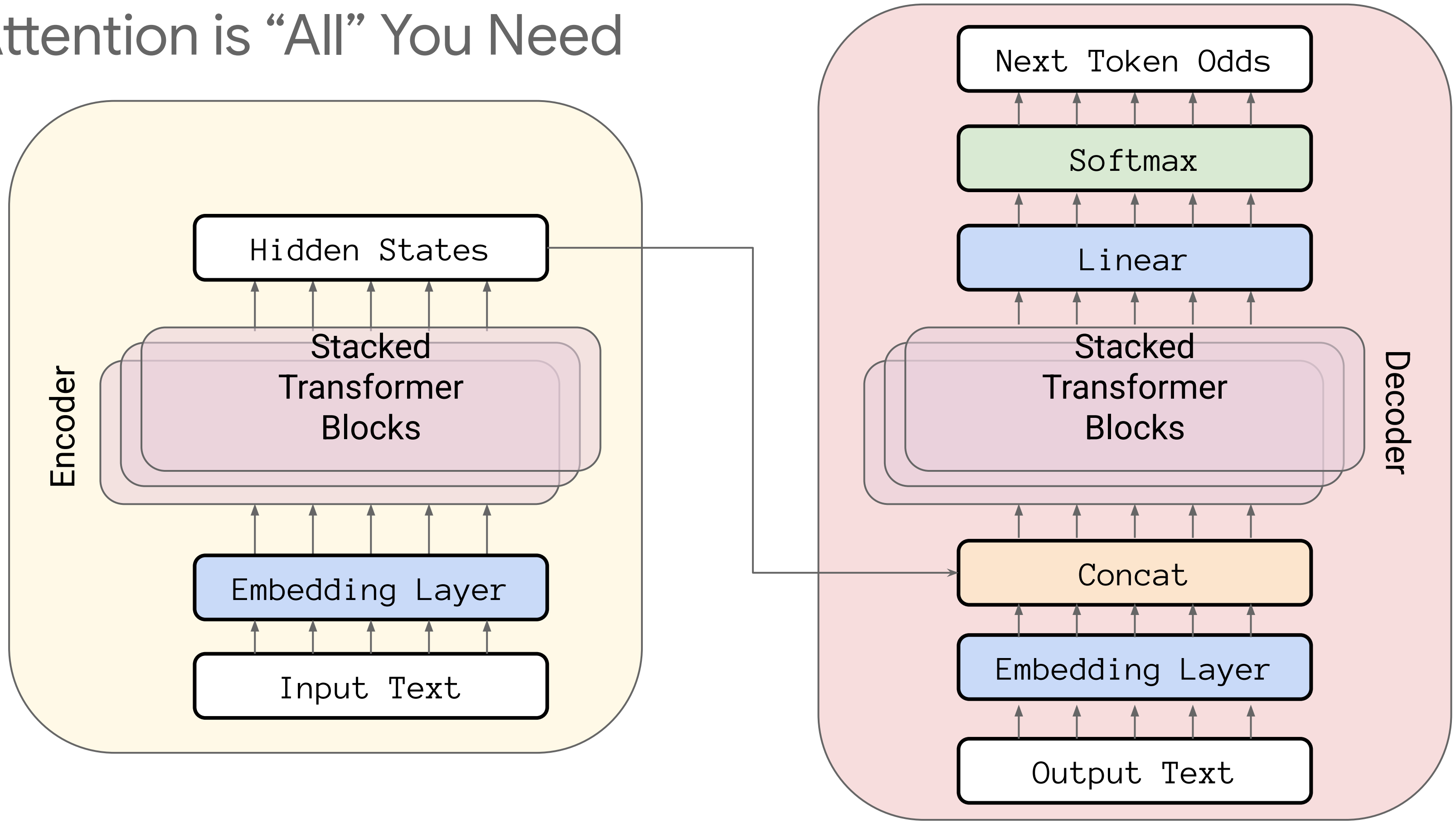
CS 4644 / 7643: Deep Learning

William Held
School of Interactive Computing
Georgia Institute of Technology

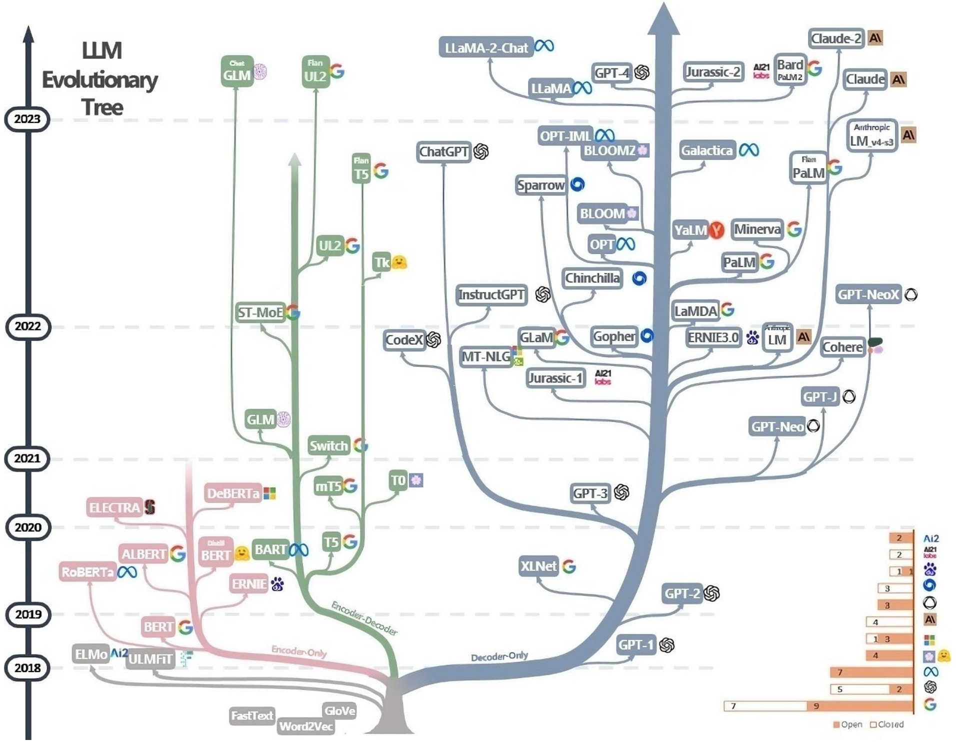
Transformer Lecture Speed Recap: The Transformer Block



Attention is “All” You Need



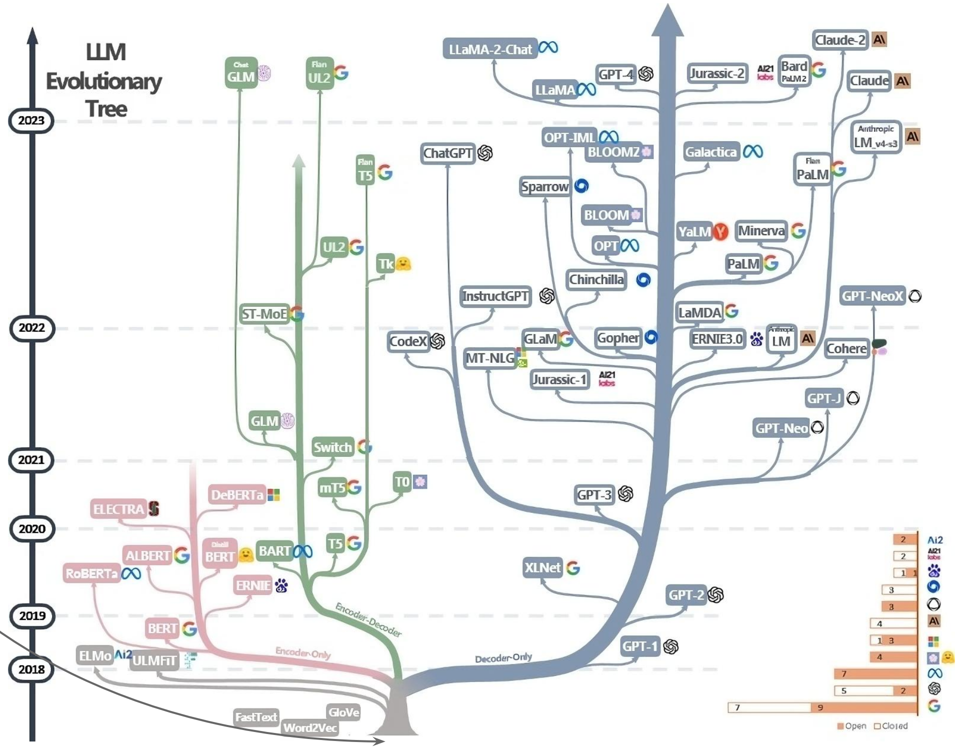
How do we go from purpose driven models to LLMs?



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

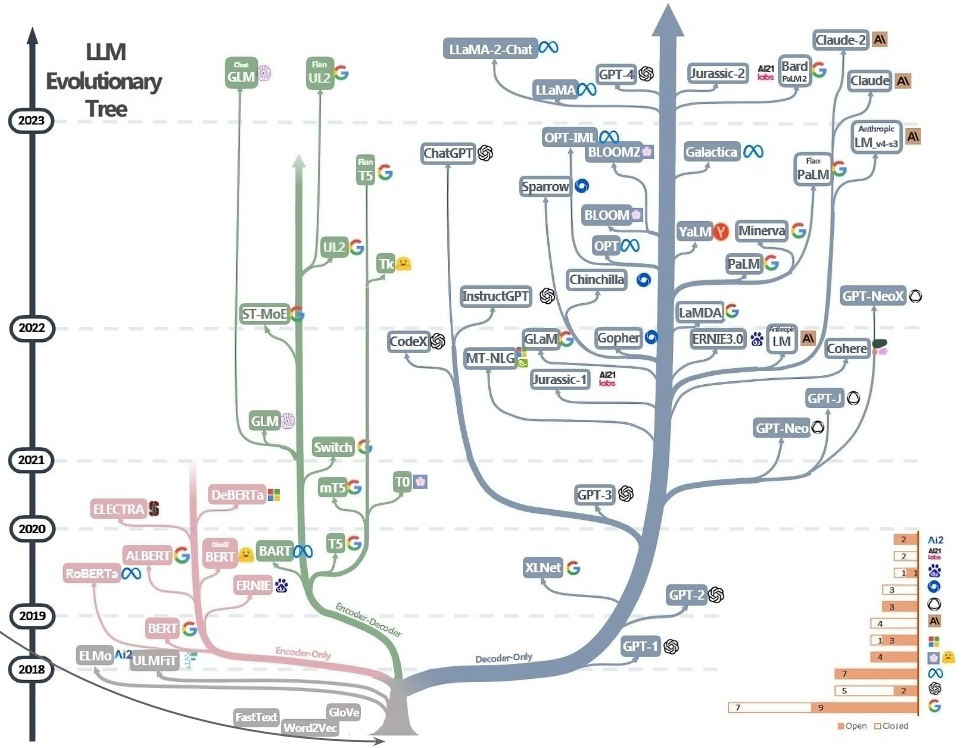
How do we go from purpose driven models to LLMs?

Self-Supervised Learning

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

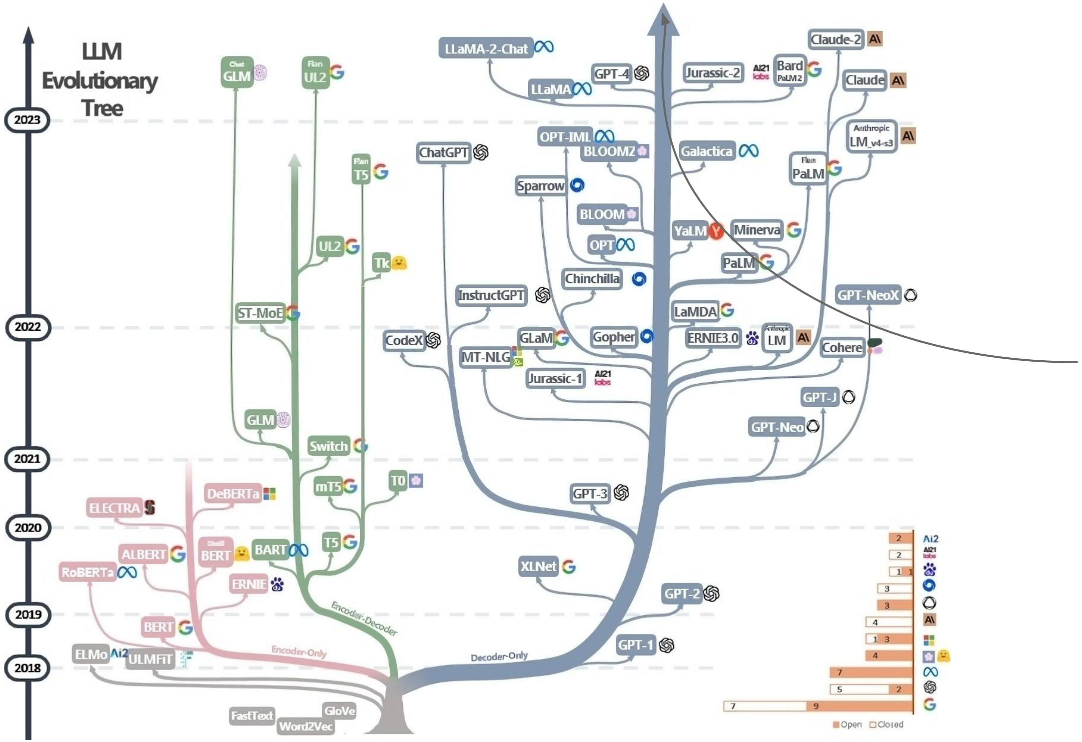
<u>Covered Today</u>

- Encoder Only
- Decoder Only



https://github.com/Mooler0410/LLMsPracticalGuide

How do we go from purpose driven models to LLMs?

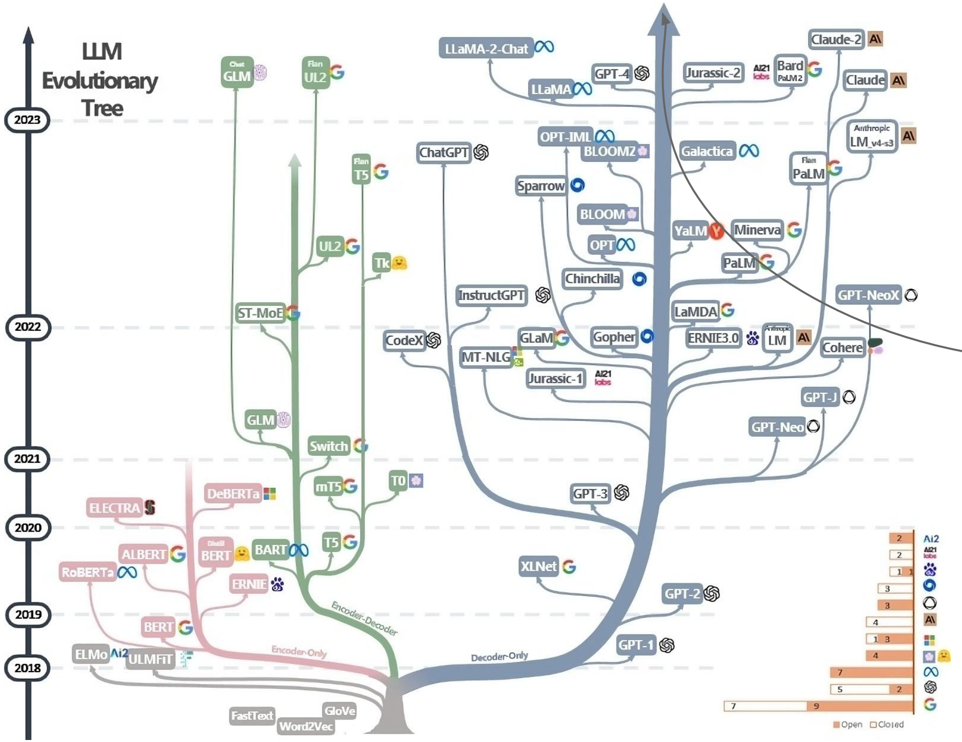


Scaling Laws

How do we train large models on large amounts of quality data?

https://github.com/Mooler0410/LLMsPracticalGuide

How do we go from purpose driven models to LLMs?



Scaling Laws

How do we train large models on large amounts of quality data?

<u>Covered Today</u>

- Data Curation
- Distributed Training
- “Alignment”

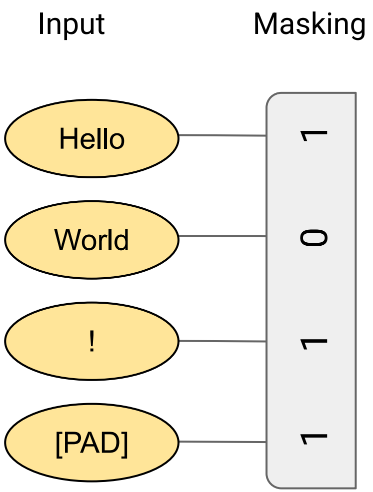
https://github.com/Mooler0410/LLMsPracticalGuide

LLM Advancements have been driven primarily by these two

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

Scaling Laws
How do we train large models on large amounts of quality data?

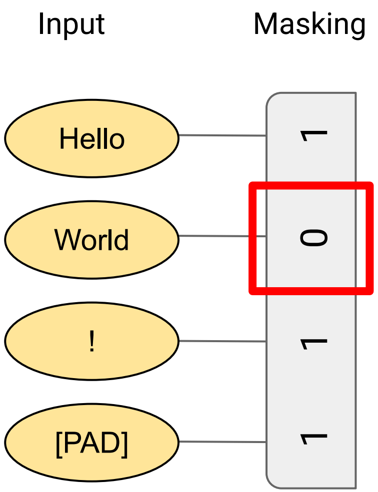
SSL | From raw text to loss!



Masked Language Model

Devlin et al. 2018 (BERT)

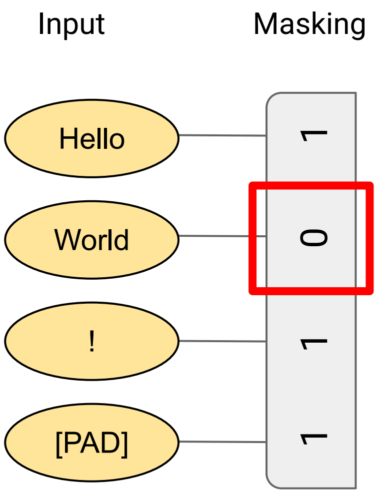
SSL | What is the “Mask” in a Masked Language Model?



Masked Language Model

Devlin et al. 2018 (BERT)

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



Masked Attention

Similarities: $E = (QXT / \mathrm{sqrt}(DQ)) * MASK$

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

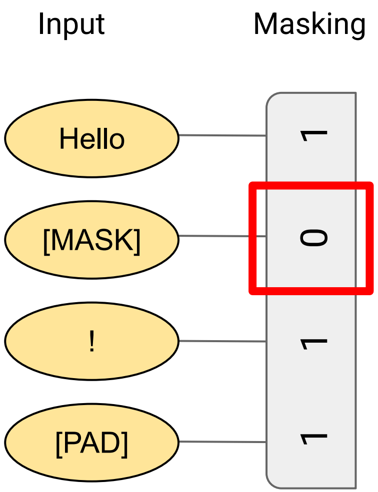
Output vectors: $Y = AX$

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

Masked Language Model

Devlin et al. 2018 (BERT)

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



Intuition

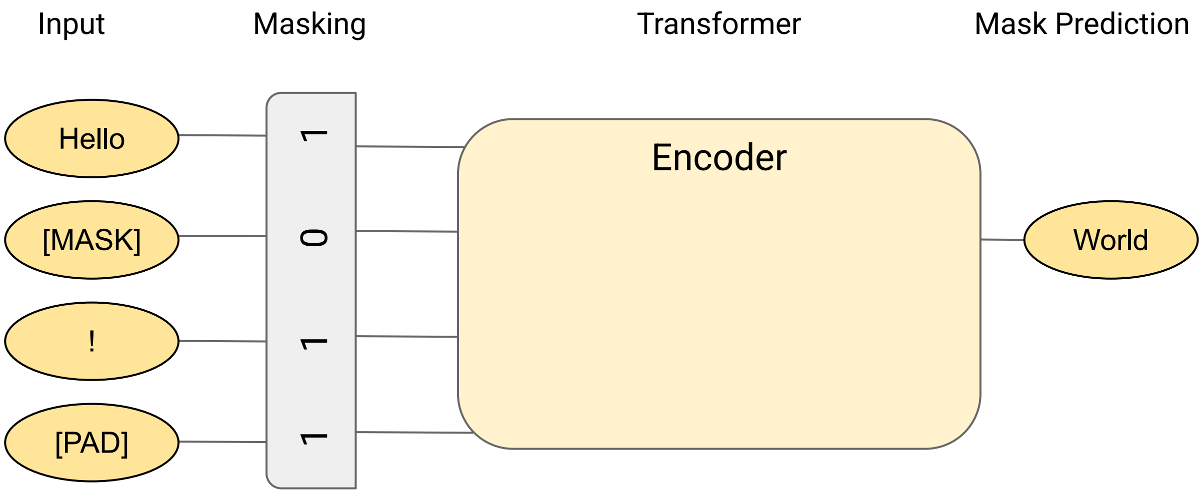
If $MASK_i = 0$, then $Y_i = \sum_{j, j != i} A_{i,j} X$

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



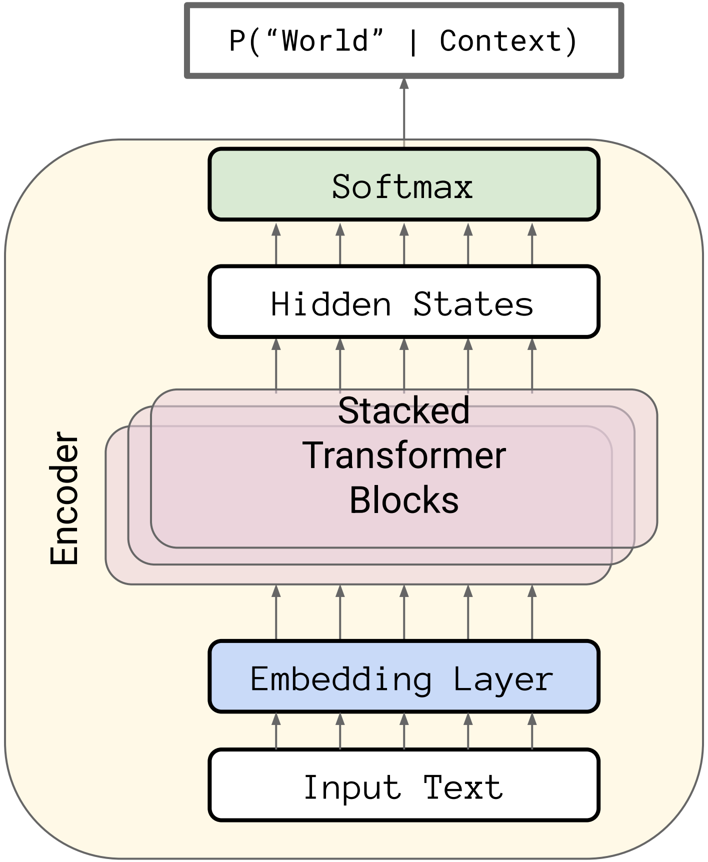
Masked Language Model

[Devlin et al. 2018 (BERT)]

SSL | Masked Token Prediction

Optimize Negative Log Likelihood

`loss = -log(P(“World” | Context)`

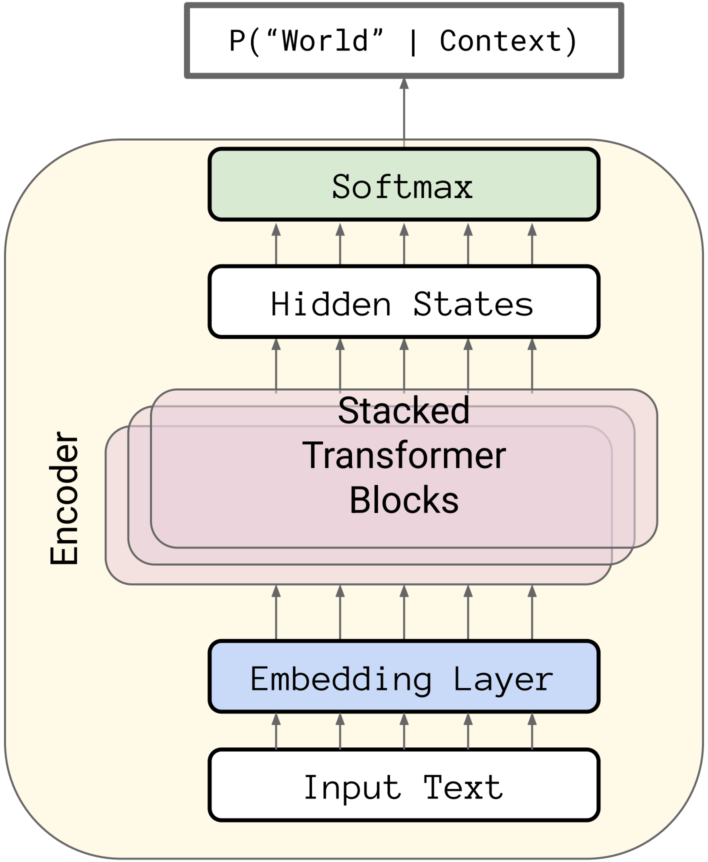


SSL | Masked Token Prediction

Optimize Negative Log Likelihood

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

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



Data | BERT used existing curation!

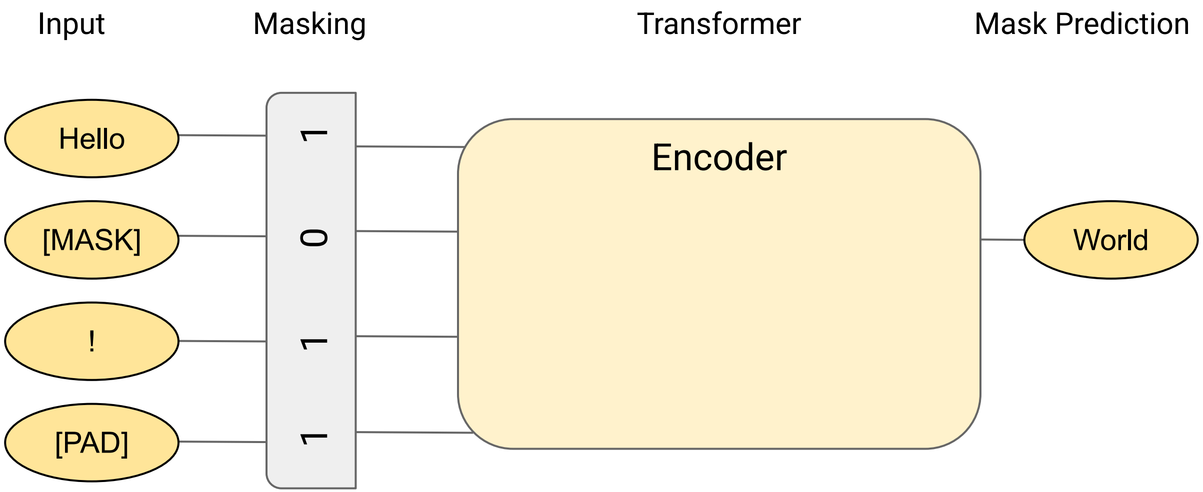
BERT Corpus
English Wikipedia + BooksCorpus

Size
~3 Billion Tokens

Quality
High quality text,
Broad “Academic” Knowledge,
Limited Diversity

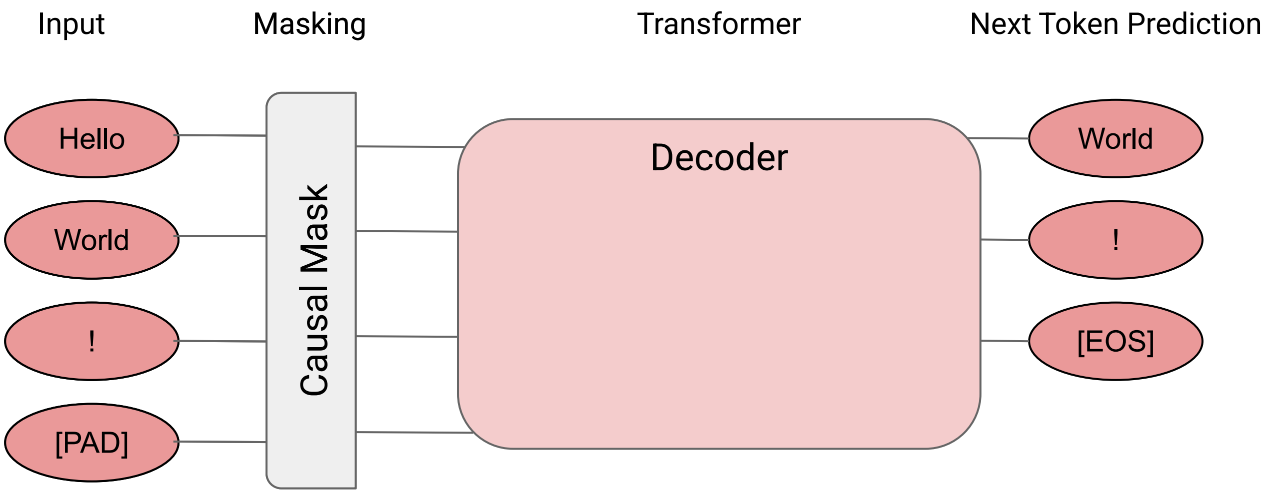
[Devlin et al. 2018 (BERT)]

Questions?



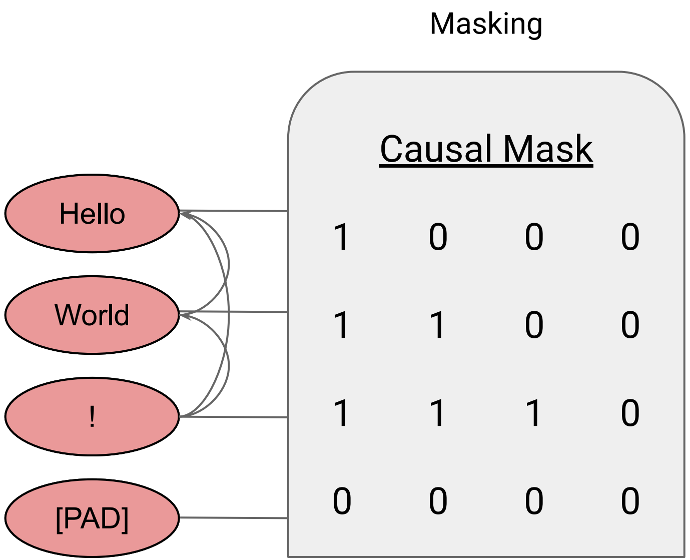
Masked Language Model

SSL | "How does GPT work?"



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

SSL | Autoregressive Language Modeling



Masked Attention Again!

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

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

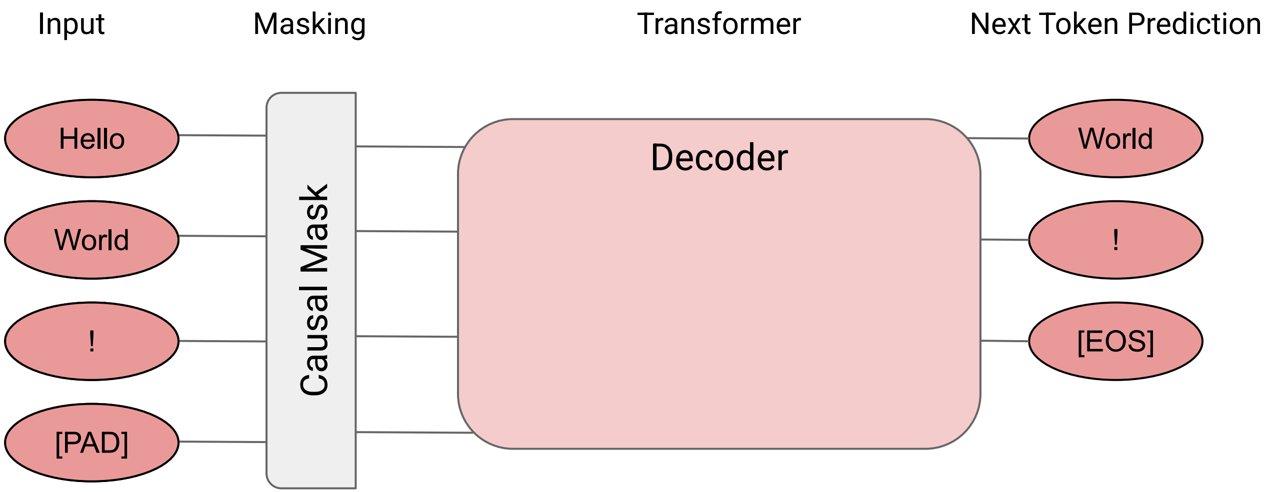
Output vectors: Y = AX

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

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

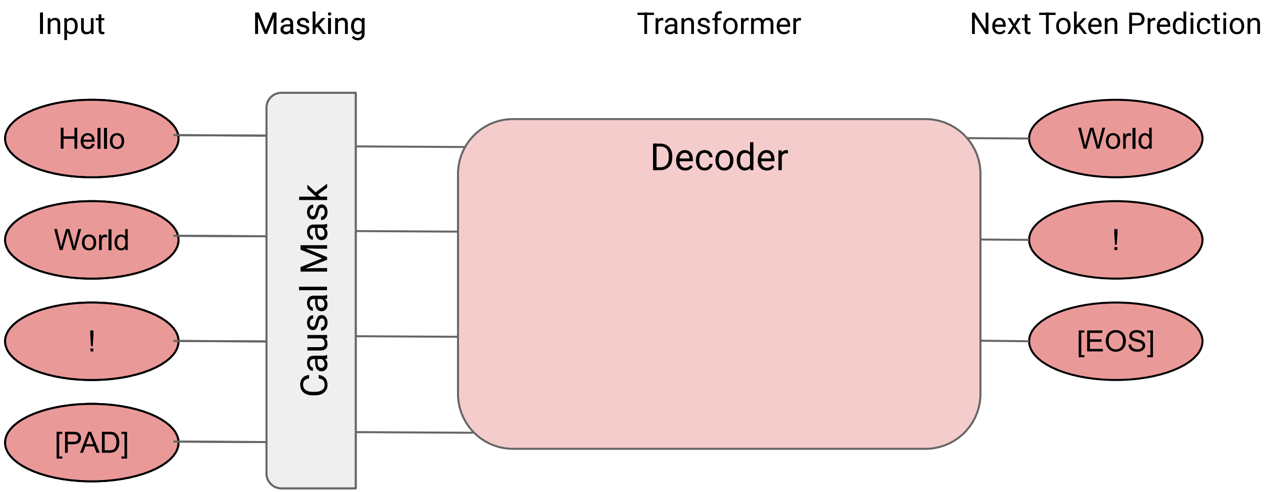
Size
~10 Billion Tokens

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)

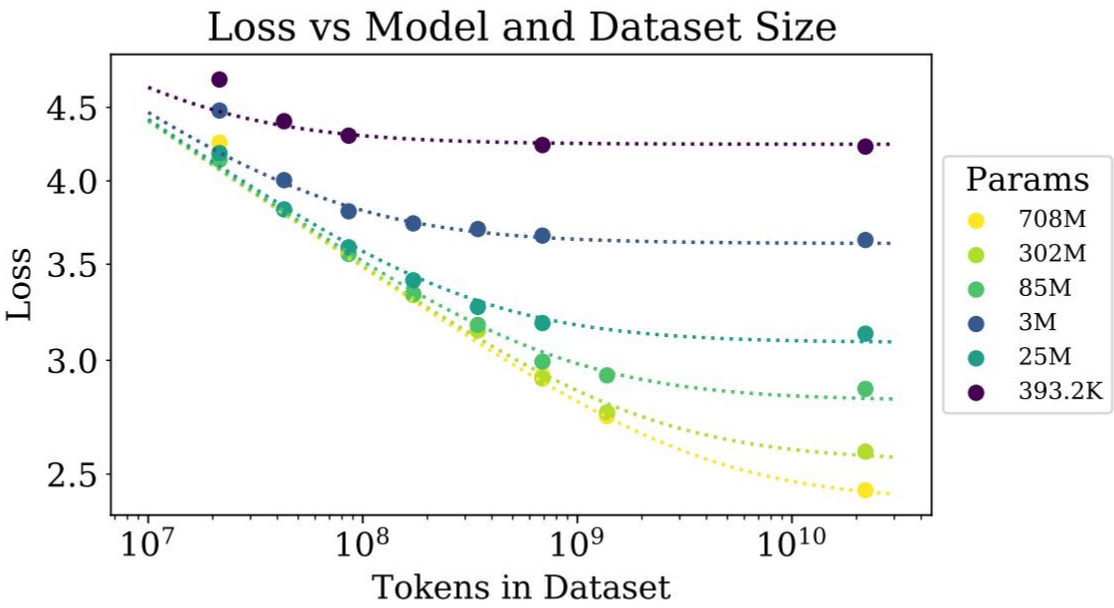
Questions?



Autoregressive Language Model

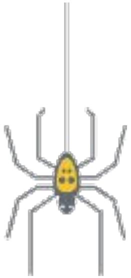
Data & Parameter Scaling | Moving to **Large** Language Models

Today's LLMs are driven by data and model scaling



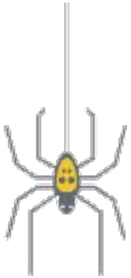
Kaplan et al. 2020

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



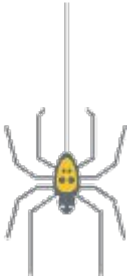
We could get a lot more data from CommonCrawl!

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

<u>T5 Corpus (AKA C4)</u>
All Common Crawl Text Which Meets Heuristics

<u>Size</u>
~350 Billion Tokens

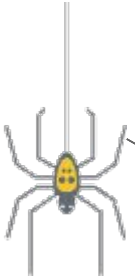
<u>Quality</u>
Varying quality text, Broad Knowledge, Improved Diversity

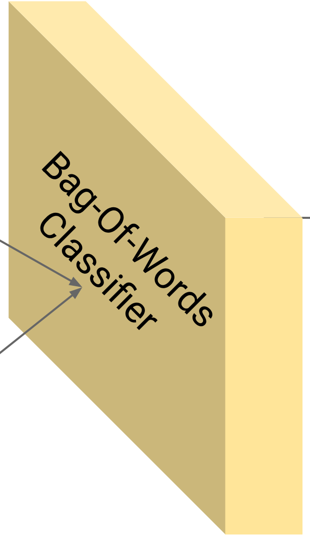
- We only retained lines that ended in a terminal punctuation mark (i.e. a period, exclamation mark, question mark, or end quotation mark).
- We discarded any page with fewer than 3 sentences and only retained lines that contained at least 5 words.
- We removed any page that contained any word on the "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

Training





Distinguish High and Low Quality

Low-Quality, High Volume

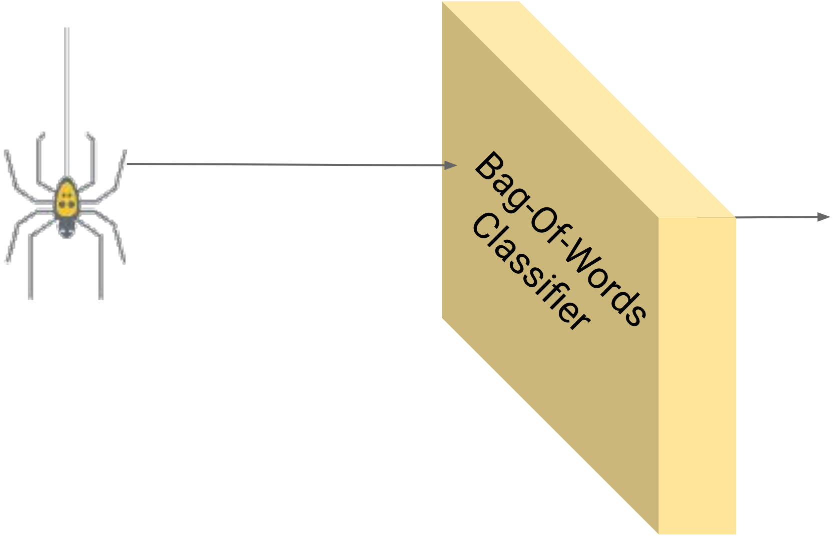
URL Domain	# Docs	% of Total Docs
bbc.co.uk	116K	1.50%
theguardian.com	115K	1.50%
washingtonpost.com	89K	1.20%
nytimes.com	88K	1.10%
reuters.com	79K	1.10%
huffingtonpost.com	72K	0.96%
cnn.com	70K	0.93%
cbc.ca	67K	0.89%
dailymail.co.uk	58K	0.77%
go.com	48K	0.63%

High Quality, Medium Volume

Brown et al. 2020

GPT-3 - Increased Scaling Via Automated Data Curation

Filtering



Keep “False” Positives

“False” positive ~= High Quality

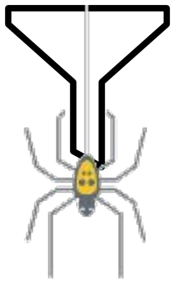
[Brown et al. 2020]

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 3 Corpus

Size
15 Trillion Tokens

Quality
Minimal details known

Dubey et al. 2024

Gemini Corpus

Size
Unknown

Quality
No details known

Anil et al. 2023

GPT-4 Corpus

Size
Unknown (Est. >11T Tokens)

Quality
No details known

OpenAI 2023

Questions?

Llama 3 Corpus

Size
15 Trillion Tokens

Quality
Minimal details known

Dubey et al. 2024

Gemini Corpus

Size
Unknown

Quality
No details known

Anil et al. 2023

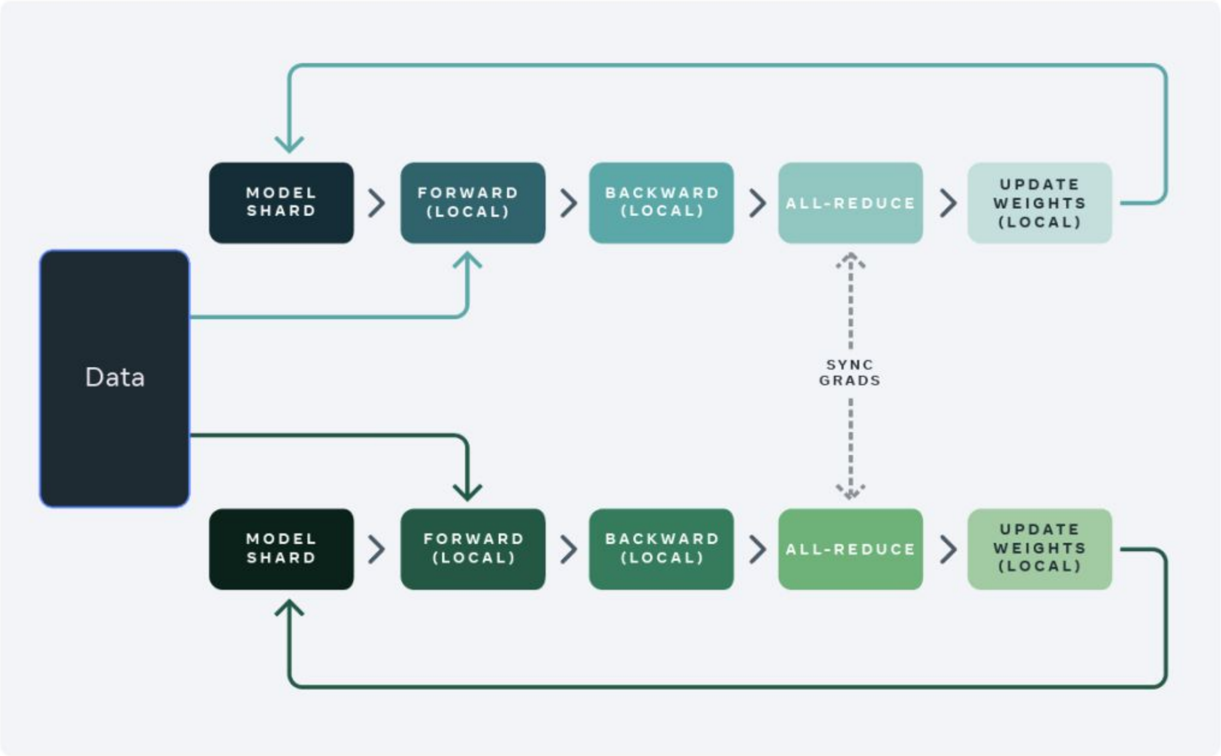
GPT-4 Corpus

Size
Unknown (Est. >11T Tokens)

Quality
No details known

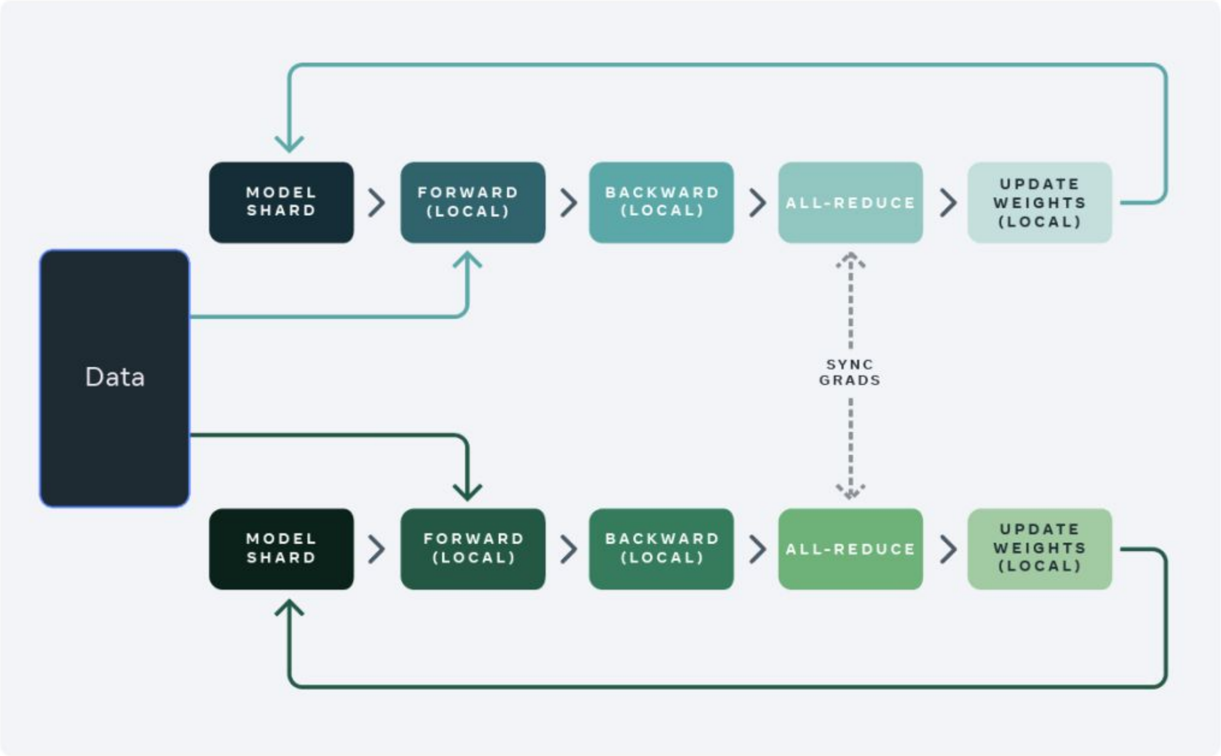
OpenAI 2023

Scaling Parameters | Data Parallel Training



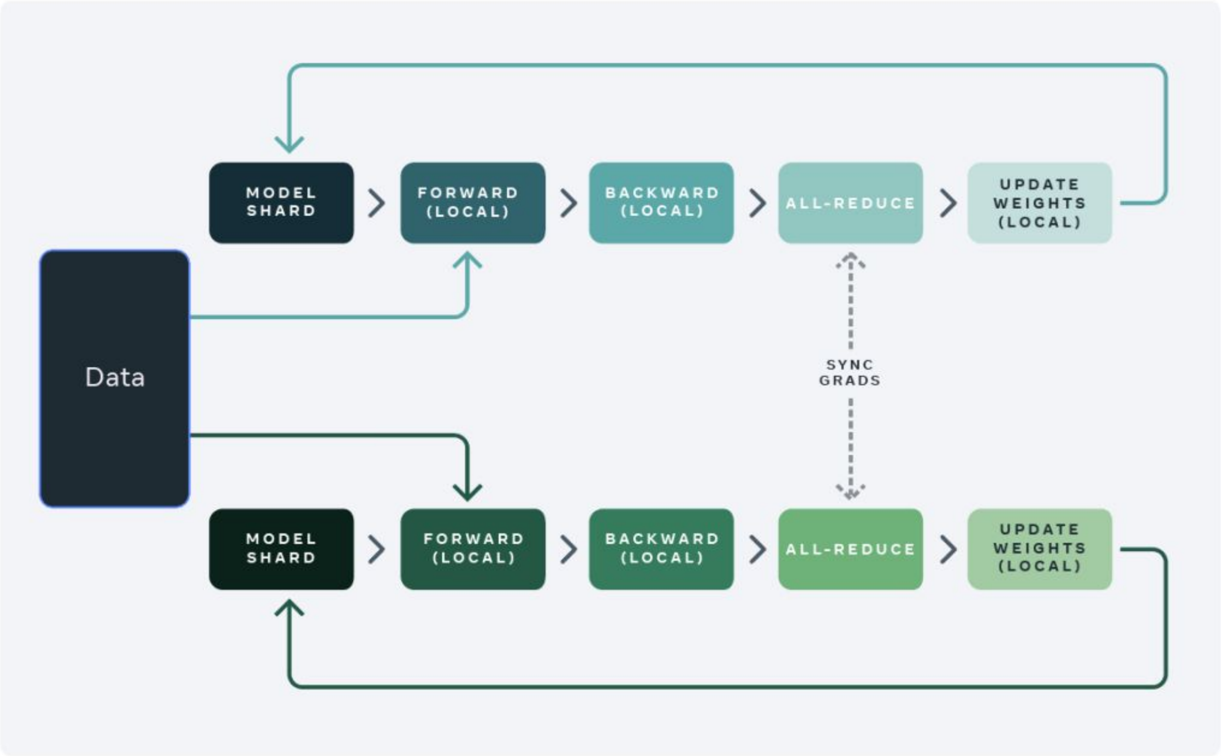
https://engineering.fb.com/2021/07/15/open-source/fsdp/

Scaling Parameters | Data Parallel Training



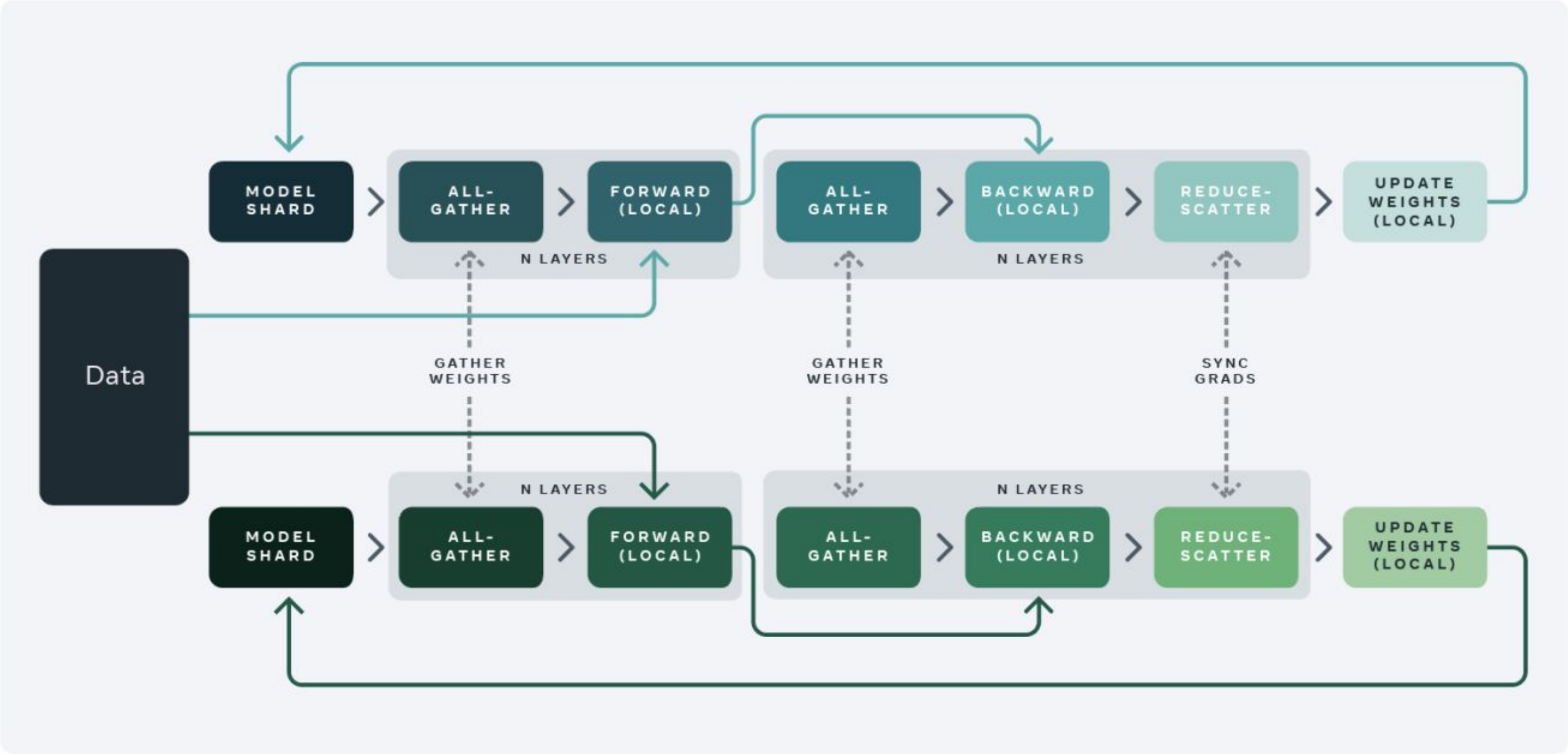
Total memory increases linearly with shards

Scaling Parameters | Data Parallel Training



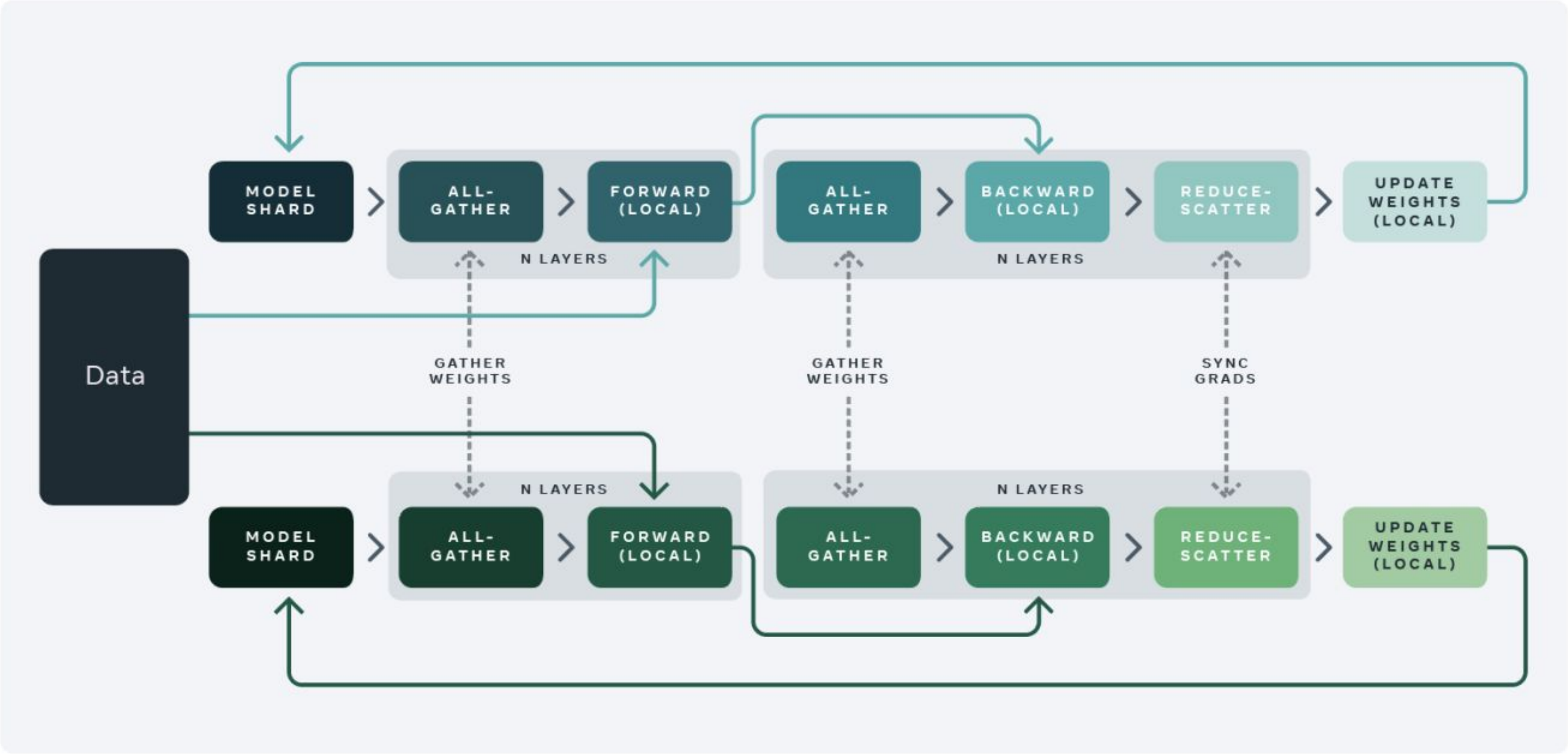
Max memory constrains model size

Scaling Parameters | *Fully* Sharded Data Parallel Training



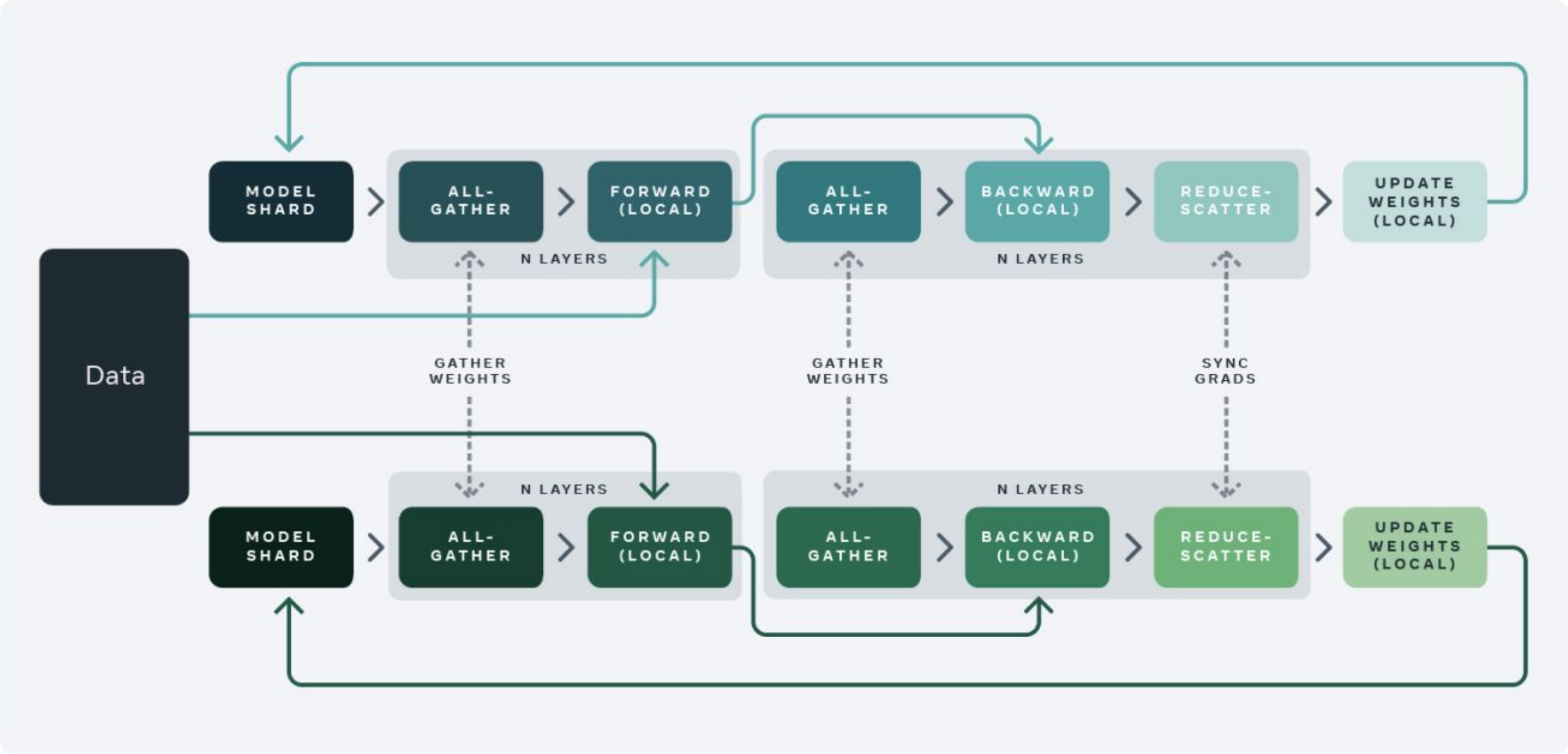
https://engineering.fb.com/2021/07/15/open-source/fsdp/

Scaling Parameters | *Fully* Sharded Data Parallel Training



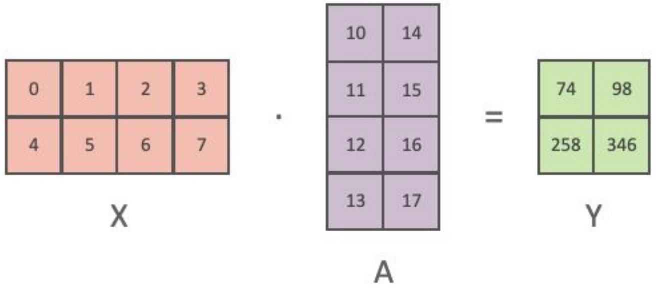
Total memory is constant

Scaling Parameters | *Fully* Sharded Data Parallel Training



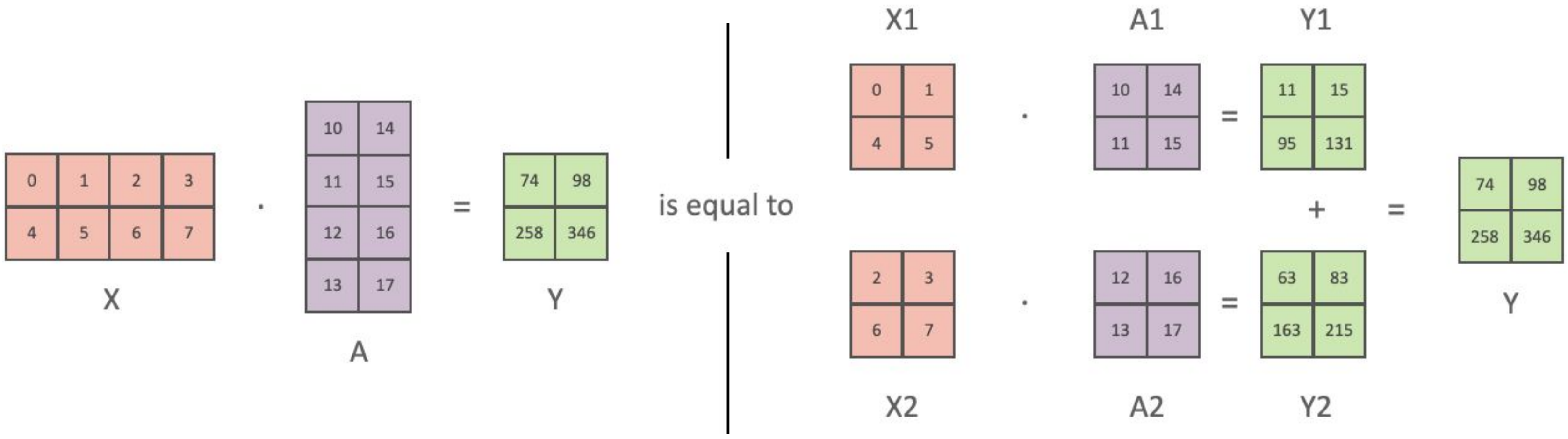
Max single GPU memory constrains layer size

Scaling Parameters | Tensor Parallel Training



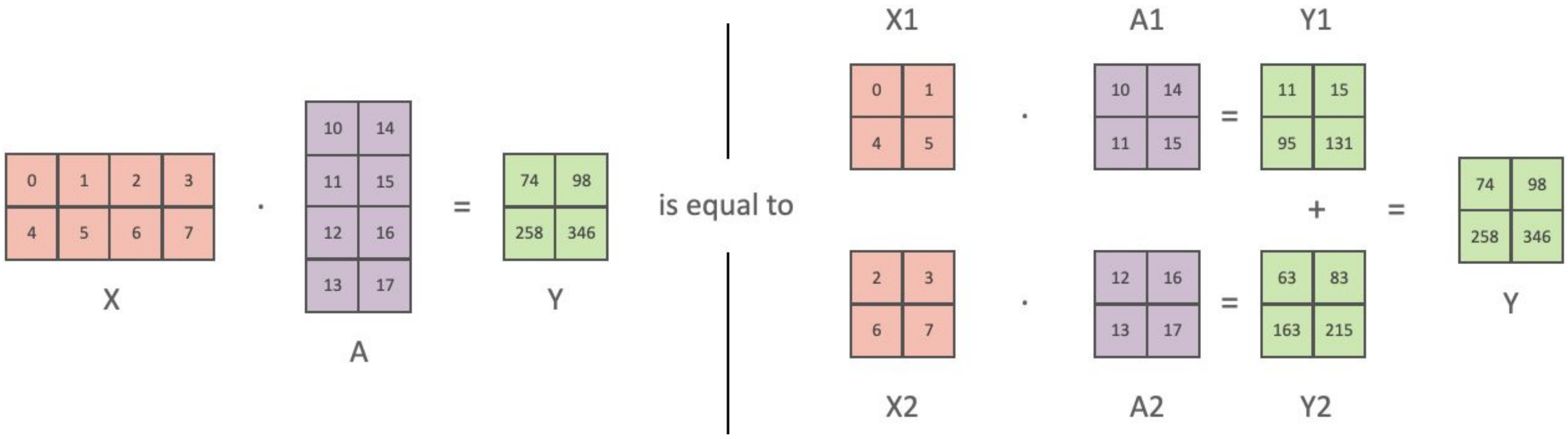
https://huggingface.co/docs/transformers/v4.15.0/parallelism#tensor-parallelism

Scaling Parameters | Tensor Parallel Training



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

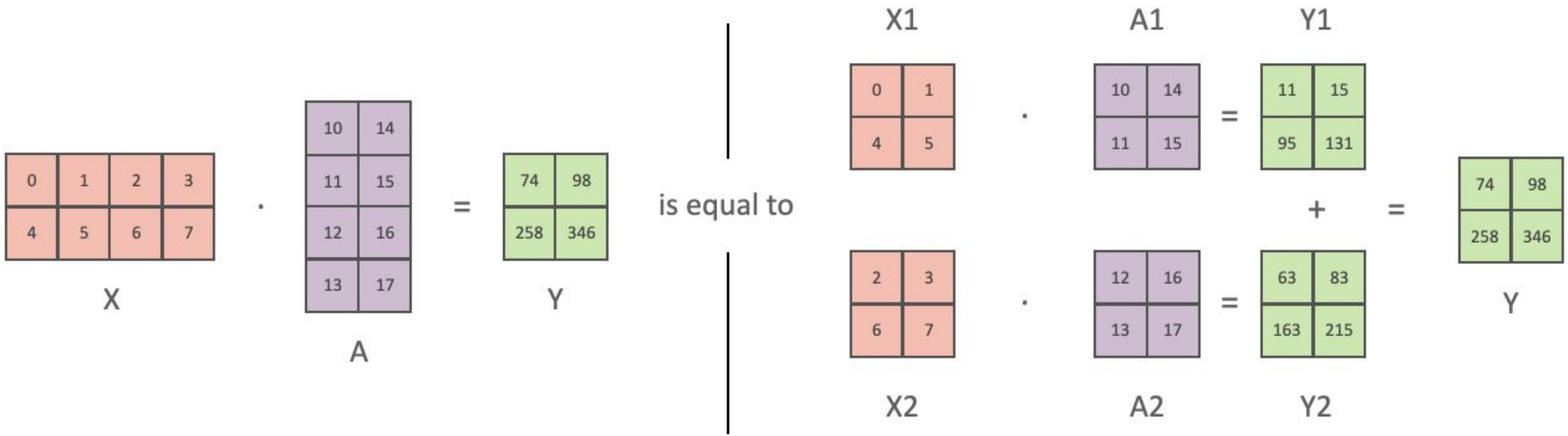
Scaling Parameters | Tensor Parallel Training



Don't need to sync gradients!

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

Scaling Parameters | Tensor Parallel Training

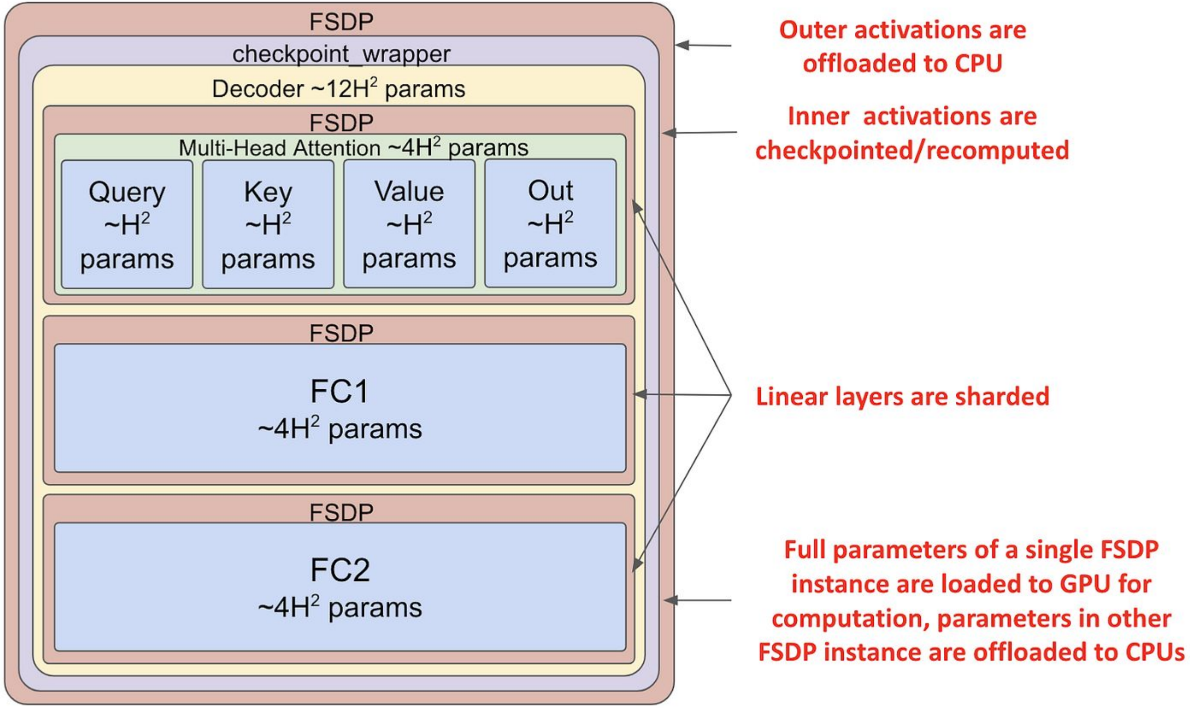


Don't need to sync gradients!

Max GPU memory constrains layer shard size

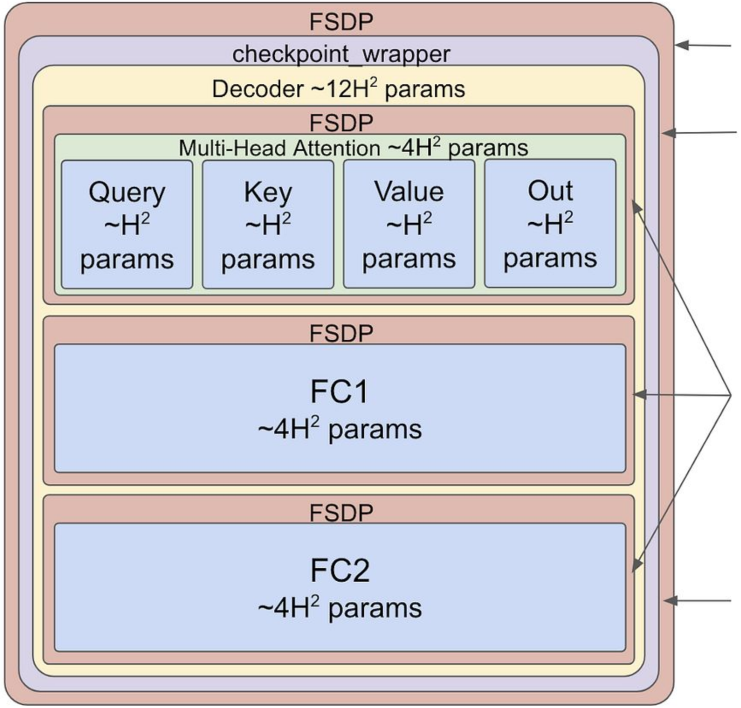
https://huggingface.co/docs/transformers/v4.15.0/parallelism#tensor-parallelism

Scaling Parameters | FSDP + TP = ~Limitless Scaling



[1 Trillion Parameter Model with Tensor Parallelism and FSDP]

Scaling Parameters | FSDP + TP = ~Limitless Scaling



Outer activations are offloaded to CPU

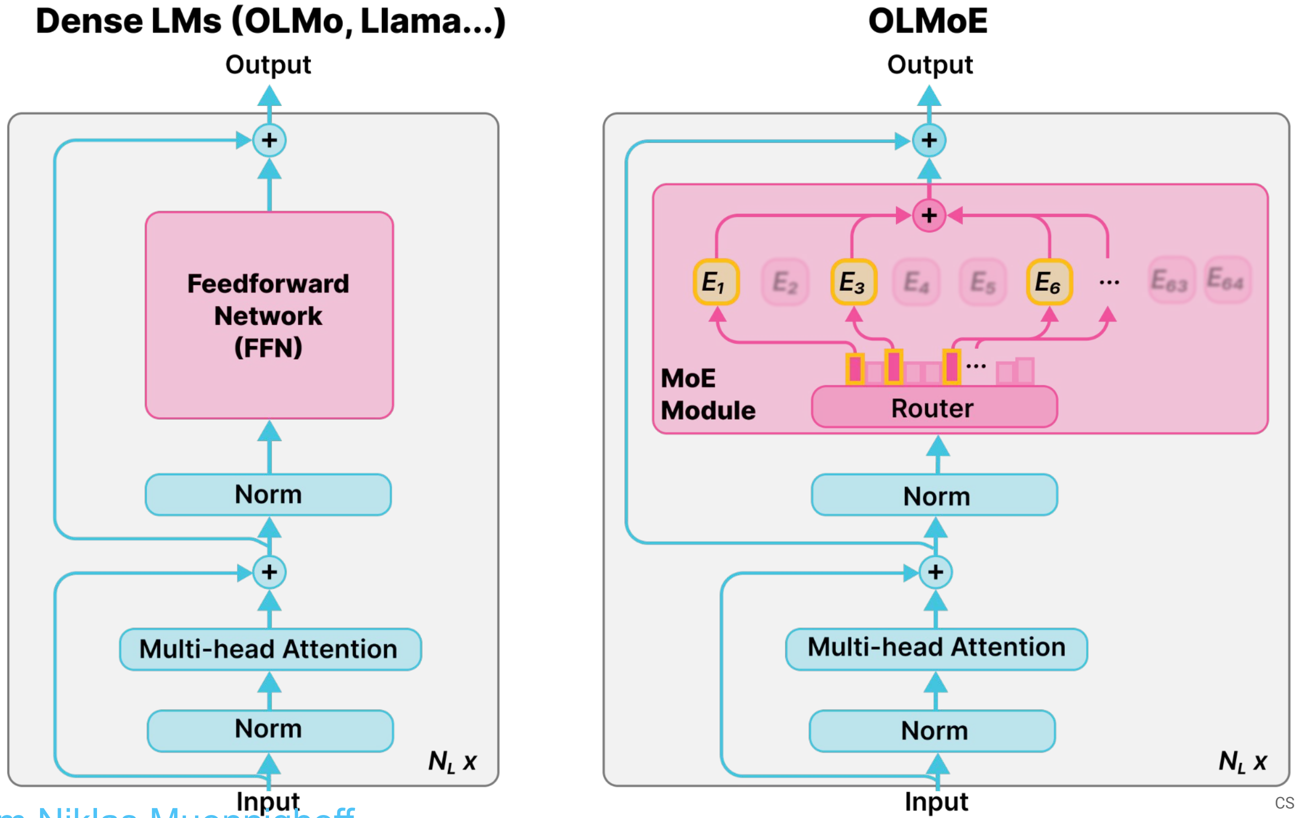
Inner activations are checkpointed/recomputed

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

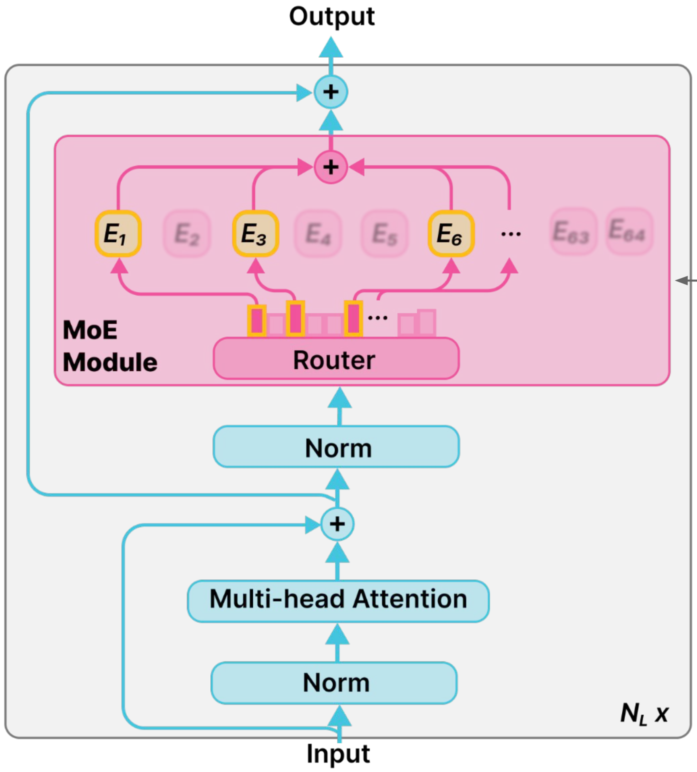
[1 Trillion Parameter Model with Tensor Parallelism and FSDP]

Scaling Parameters | Mixture Of Experts



[Slides adapted from Niklas Muennighoff]()

Scaling Parameters | Mixture Of Experts



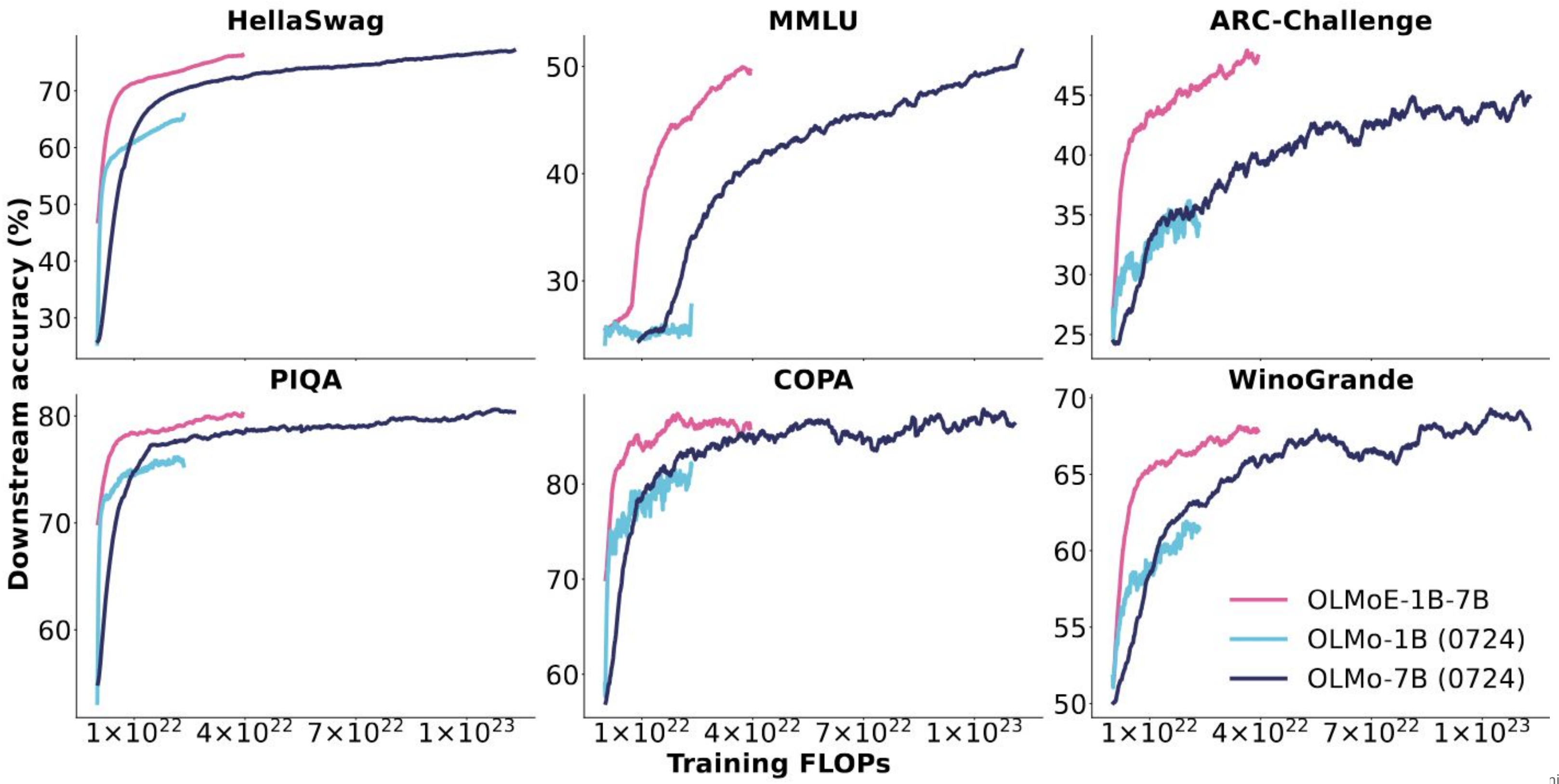
Router activation is sparse!

Pros of Mixture of Experts

- \+ Cheaper, Large Scale Training
- \+ Lower Inference Requirements

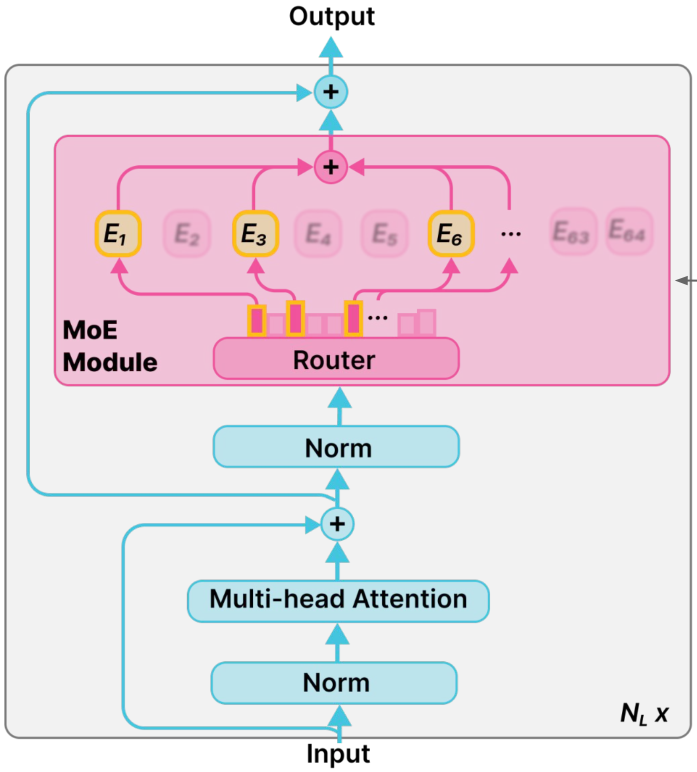
[Slides adapted from Niklas Muennighoff]

Scaling Parameters | Mixture Of Experts



[Slides adapted from Niklas Muennighoff]

Scaling Parameters | Mixture Of Experts



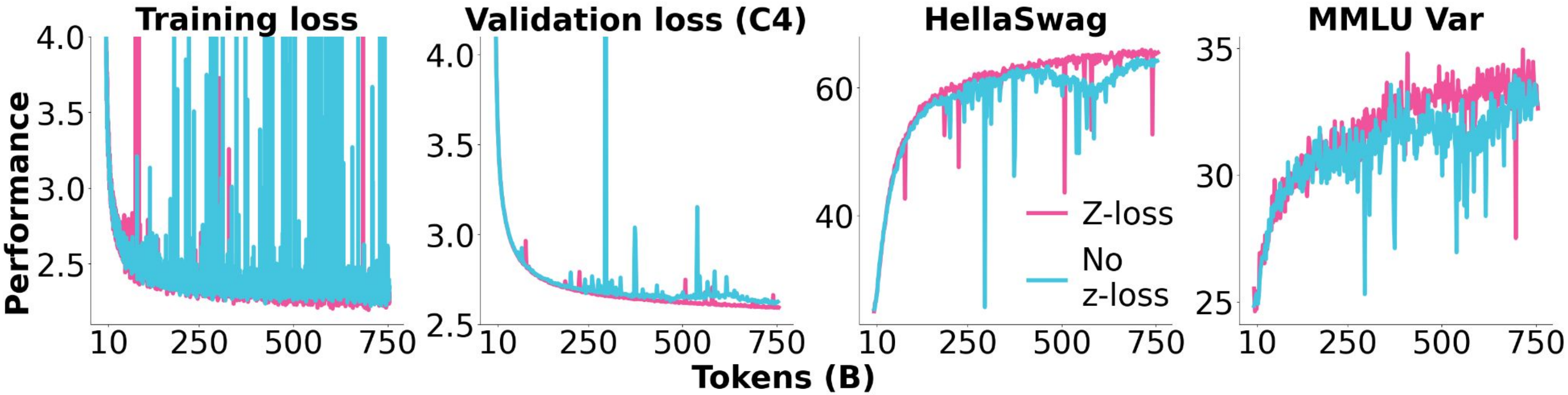
Router activation is sparse!

<u>Cons of Mixture of Experts</u>

- More Unstable Training Runs

[Slides adapted from Niklas Muennighoff]

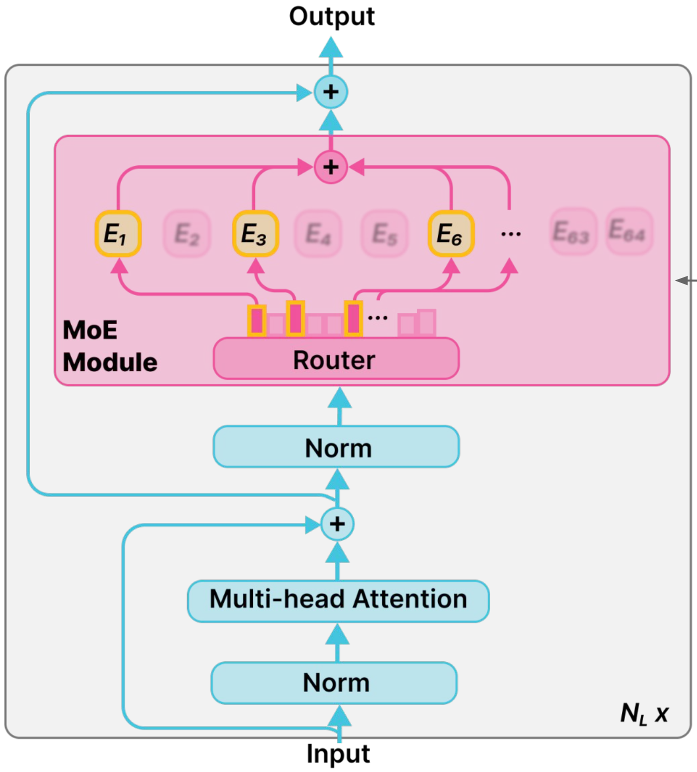
Scaling Parameters | Mixture Of Experts



Z-Loss decreases the magnitude of the logits
into the router to stabilize gradients

[Slides adapted from Niklas Muennighoff]

Scaling Parameters | Mixture Of Experts



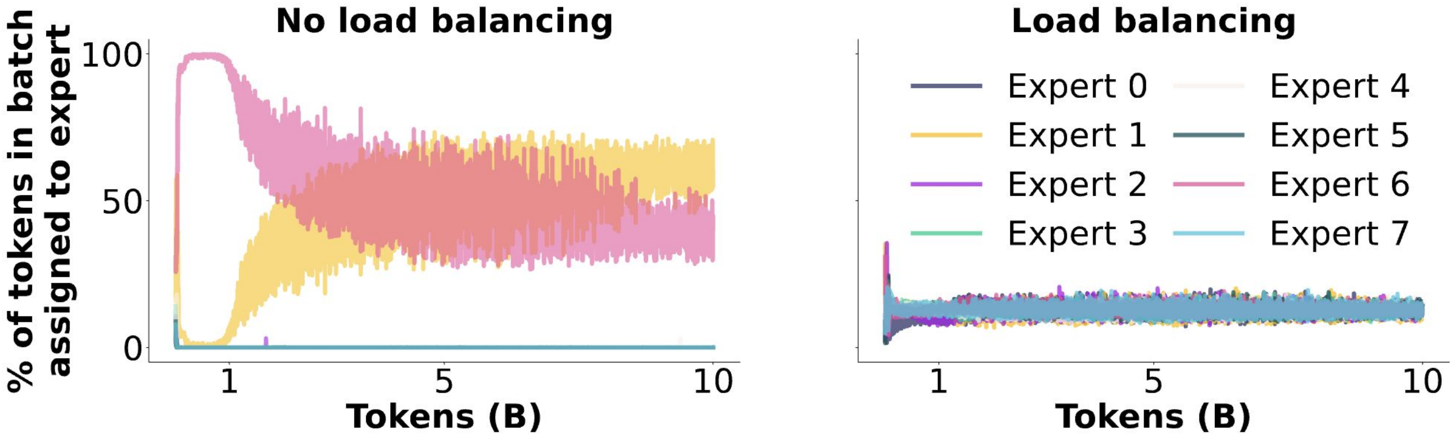
Router activation is sparse!

Cons of Mixture of Experts

- More Unstable Training Runs
- Experts frequently collapse

Slides adapted from Niklas Muennighoff

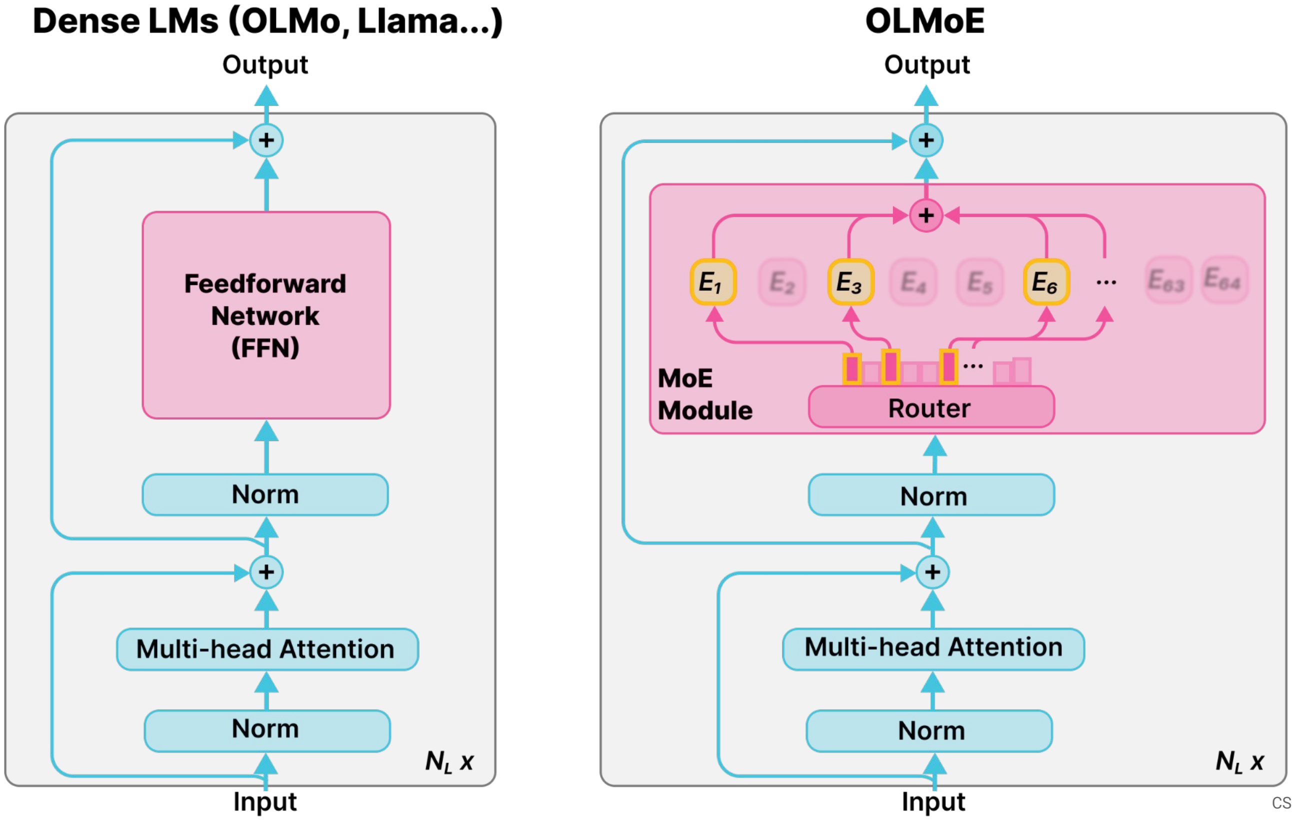
Scaling Parameters | Mixture Of Experts



Load Balancing loss to update all experts similarly

[Slides adapted from Niklas Muennighoff]

Questions?


Dense LMs (OLMo, Llama...)
Output
Feedforward Network (FFN)
Norm
Multi-head Attention
Norm
N_L x
Input
OLMoE
Output
E_1
E_2
E_3
E_4
E_5
E_6
...
E_{63}
E_{64}
MoE Module
Router
Norm
Multi-head Attention
Norm
N_L x
Input

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.

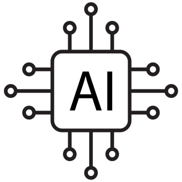
Adapted from Content by Michael Ryan @michaelryan207 for https://arxiv.org/abs/2402.15018

Massively Multitask Supervised “Instruction” Finetuning

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

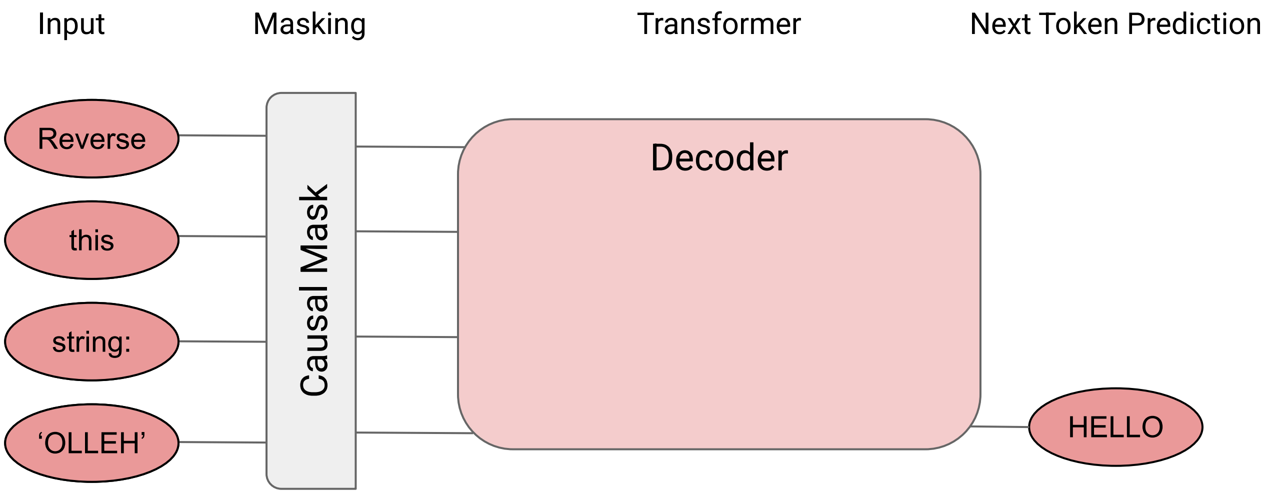
“Suggest 3 things to do in San Mateo.” + “1. Visit San Mateo, Central Park…”

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



Adapted from Content by Michael Ryan @michaelryan207 for https://arxiv.org/abs/2402.15018

Instruction Tuning | Just keep training!

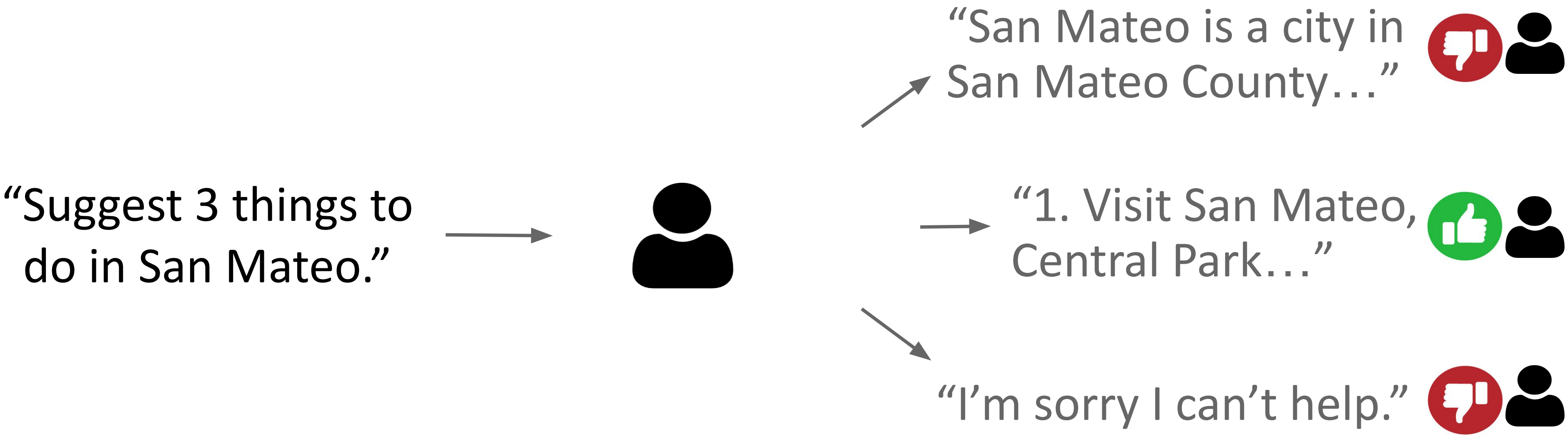


Optimize Negative Log Likelihood of The Response

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

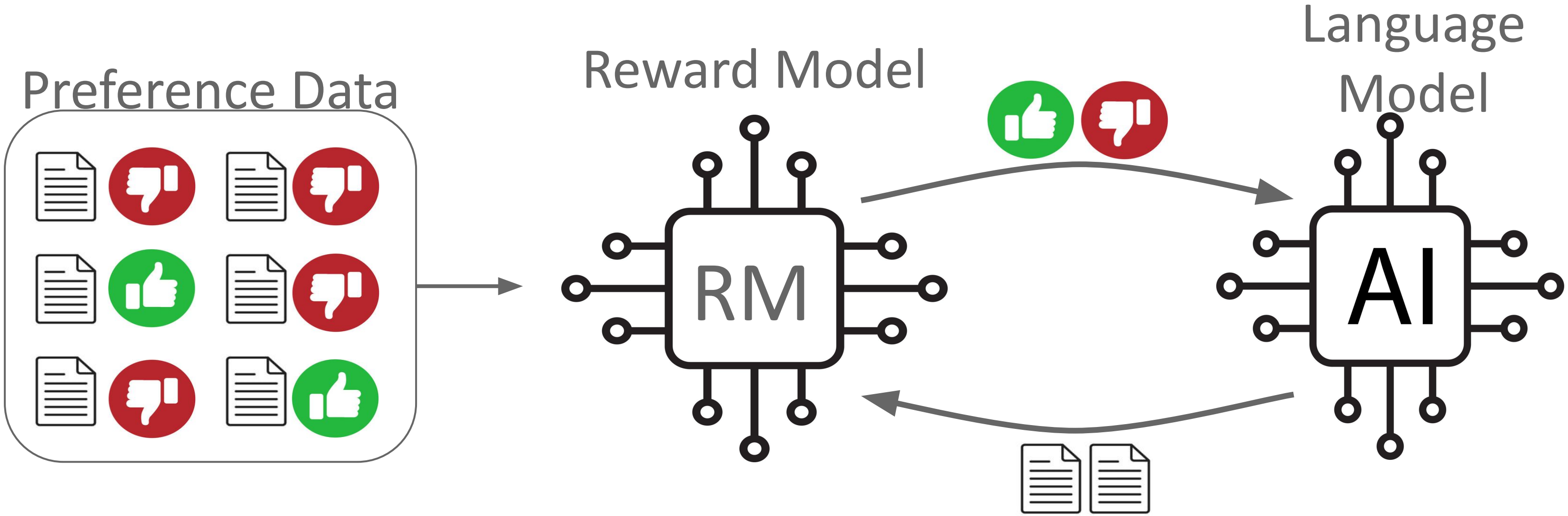
[Wang et al. 2022]

Further Refinement from Sparse Reward (RLHF)



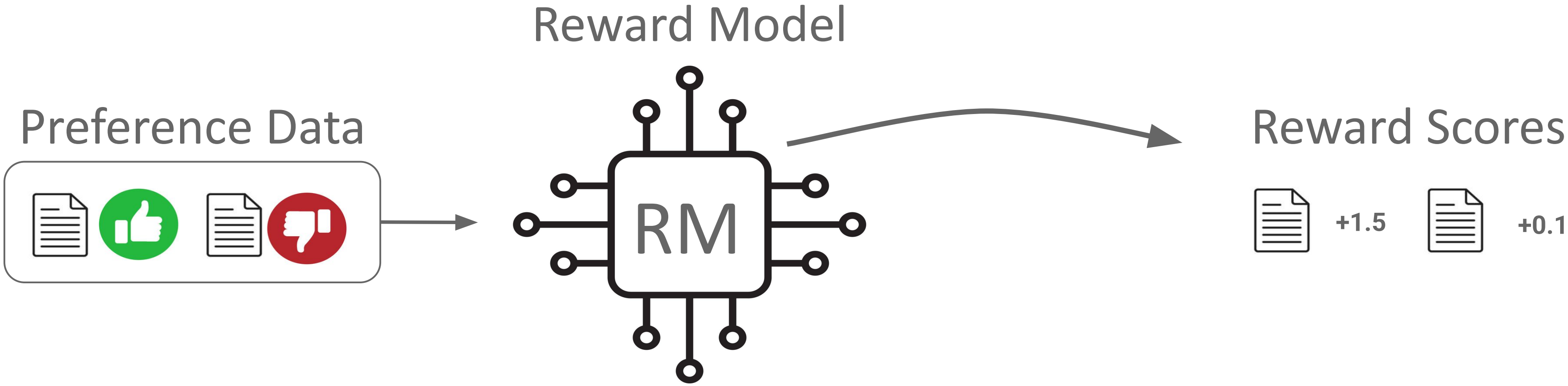
Adapted from Content by Michael Ryan @michaelryan207 for https://arxiv.org/abs/2402.15018

Reinforcement Learning From Human Feedback (RLHF)



Adapted from Content by Michael Ryan @michaelryan207 for https://arxiv.org/abs/2402.15018

Reinforcement Learning From Human Feedback (RLHF)

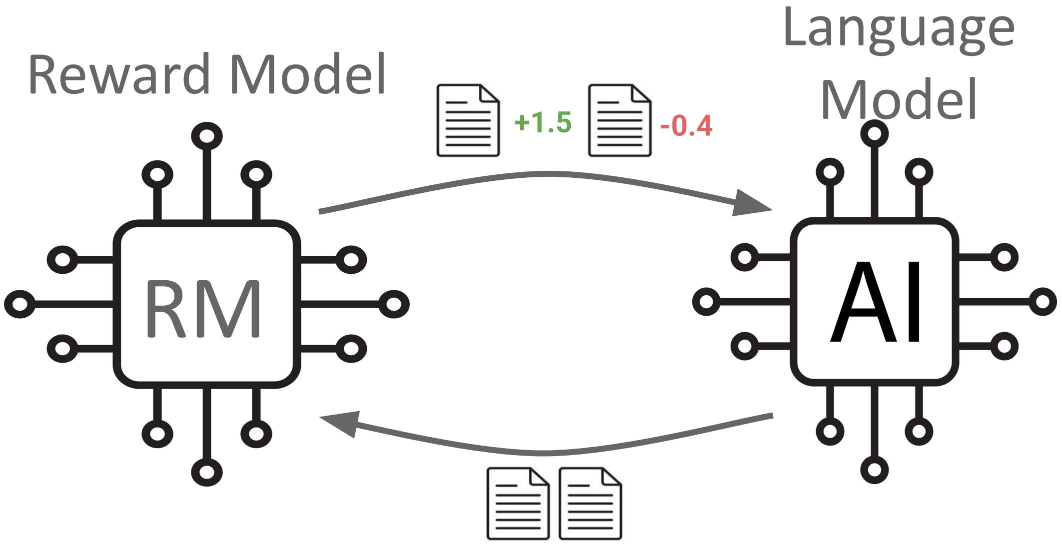


Optimize Reward Margin between Preferences

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

Adapted from Content by Michael Ryan @michaelryan207 for https://arxiv.org/abs/2402.15018

Reinforcement Learning From Human Feedback (RLHF)

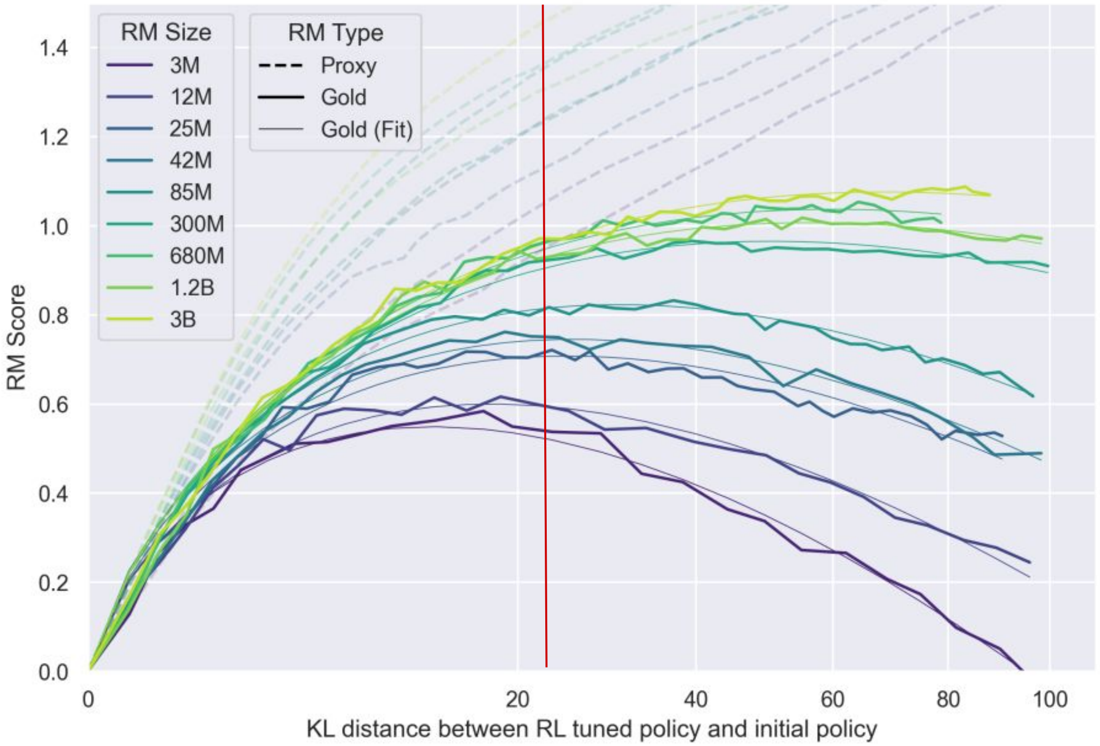


Optimize Reward Margin between Preferences

$\texttt{loss}_{\texttt{RM}}$ `= -RM(GENERATED_EXAMPLES)`

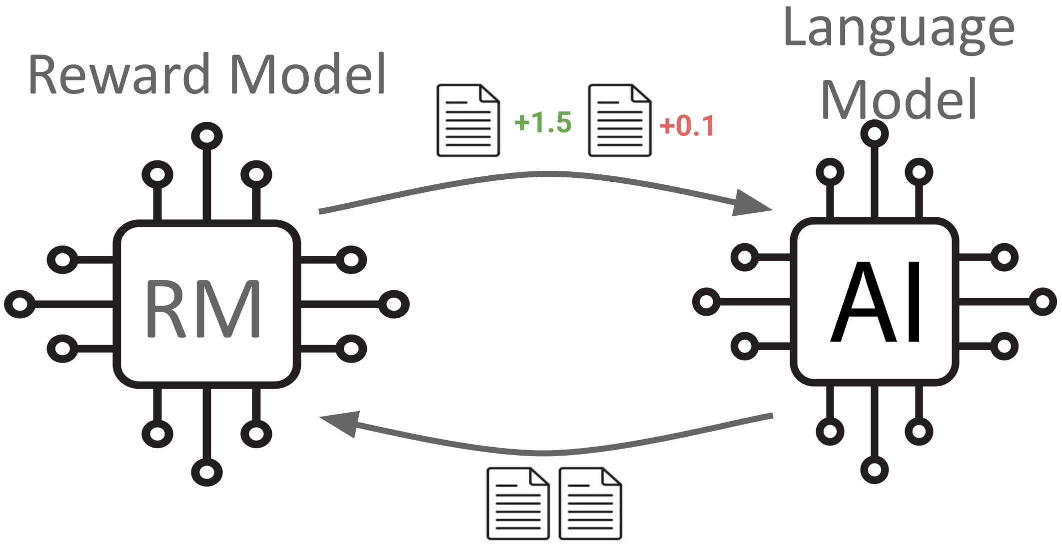
Adapted from Content by Michael Ryan @michaelryan207 for https://arxiv.org/abs/2402.15018

Models Quickly Overfit to Naively Optimized Reward



Gao et al. 2022

Reinforcement Learning From Human Feedback (RLHF)

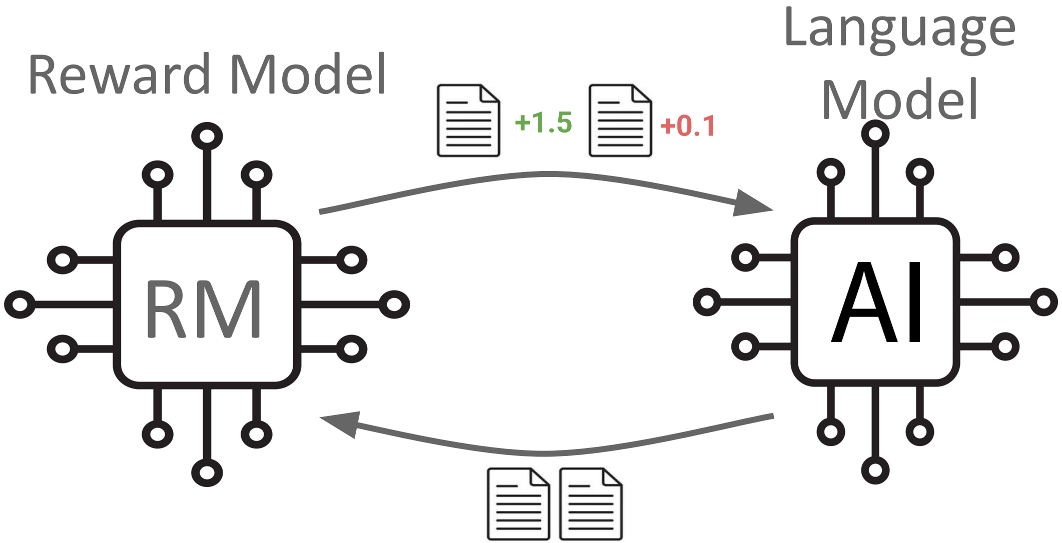


Optimize Reward Without Drifting Too Far from SFT

$$\mathtt{loss}_{\mathtt{RLHF}} = \mathtt{loss}_{\mathtt{RM}} + \mathtt{KL}(\mathtt{LM}_{\mathtt{RLHF}}, \mathtt{LM}_{\mathtt{SFT}})$$

Adapted from Content by Michael Ryan @michaelryan207 for https://arxiv.org/abs/2402.15018

Questions?



Adapted from Content by Michael Ryan @michaelryan207 for https://arxiv.org/abs/2402.15018

Final Questions?

Fill out my anonymous feedback form