

Lecture 3

EVERYTHING YOU DIDN'T WANT TO KNOW ABOUT
LM ARCHITECTURE AND TRAINING

CS336

Tatsu H

Logistics

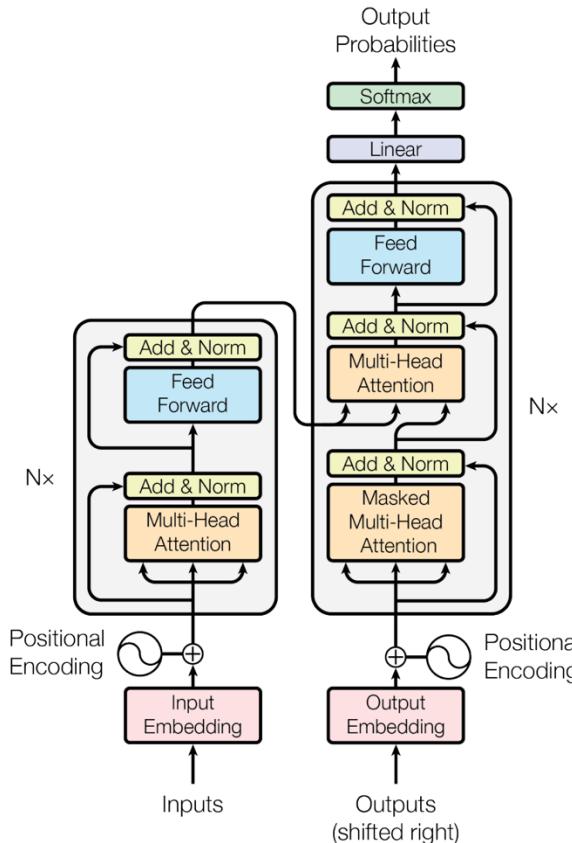
- ❖ Join the slack!
- ❖ Check to make sure you have the latest version of the assignment!

Outline and goals

- ❖ Quick recap of the ‘standard’ transformer (what you implement)
- ❖ What do most of the large LMs have in common?
- ❖ What are common variations to the architecture / training process?

Today's theme: the best way to learn is hands-on experience
the second best way is to try to learn from others' experience

Starting point: the ‘original’ transformer



Review: choices in the standard transformer

Position embedding: sines and cosines

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{\text{model}}})$$

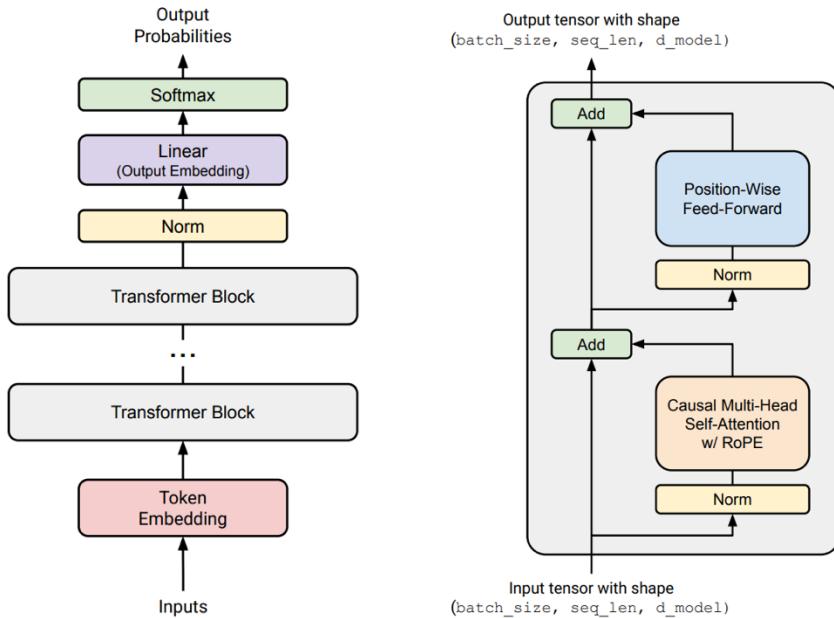
$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{\text{model}}})$$

FFN: ReLU

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Norm type: post-norm, LayerNorm

What you implemented – simple, modern variant



Differences:

- **LayerNorm** is in front of the block
- **Rotary position embeddings (RoPE)**
- FF layers use **SwiGLU**, not ReLU
- Linear layers (and layernorm) have **no bias** (constant) terms

Why did we pick these?
What should you pick?

FALCON2-11B TECHNICAL REPORT

Quentin Maliseti^{*} Nilshetho Das Chowdhury^{*}

Ruxandra Cojocaru Mugariya Farooq

i Sanath Narayan Ankit Singh

Mohammed Al-Yafeai Hamza Alobeidli

Kirill Fedyakin Reda Alami

Ihab Alnabi, United Arab Emirates

[de] [de]



The Llama 3 Herd of Models

Llama Team, AI @ Meta¹

¹A detailed contributor list can be found in the appendix of this paper.

Nemotron-4 340B Technical Report

NVIDIA

Qwen2.5 Technical Report

Qwen Team

<https://huggingface.co/Qwen>

Reka Core, Flash, and Edge: A Series of Powerful Multimodal Language Models

Aitor Ormazabal Che Zheng Cyprien de Masson d'Autume Dani Yogatama

Deyu Fu Donovan Ong Eric Chen Eugenie Lamprecht Hai Pham Isaac Ong

Kaloyan Aleksiev Lei Li Matthew Henderson Max Bain Mikel Artetxe

Nishant Relan Piotr Padlewski Qi Liu Ren Chen Samuel Phua

Yazheng Yang Yi Tay Yuqi Wang Zhongkai Zhu Zhihui Xie

2025-03-12

Over 19 new dense model releases, many of them with minor architecture tweaks..

Shane Arora^{*} Akshita Bhaoia^{*} Yulina Gu^{*} Shuo

Zhao, Jinch Bai, Junzhe He, Juyang Lin, Kai Tang, Kehui Chen, Kezhi Tang, Ming Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruizhe Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihua Fan

InternLM2 Technical Report

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Oipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tai Jiang, Penglong Jian, Zhenjian Jin, Zhikai Lei, Jiaxing Li, Jimin Lv, Yiqi Ma, Jiarui Huang, Kewen Liu, Kuikuan Liu, Xianran Lv, Chengqi Lv, Huijun Lv, Kai Lv, Li Lv, Runyuan Ma, Zezun Ma, Wenchang Ning, Linke Ouyang, Jianqiao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Wang, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Xirong Yan, Xiaogui Yang, Haochen Yu, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zhang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wen Zhang, Weiyue Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengze Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, Dahua Lin

Shanghai AI Laboratory
SenseTime Group
The Chinese University of Hong Kong
Fudan University
internlm@pjlab.org.cn

of a Small Language Model

such^{*} Gabriel Martín Blázquez^{*} Guilherme Penedo ilícek Agustín Piñeres Lajarin Vaibhav Srivastav guyen Clémantine Fourrier Ben Burtneshaw o Cyril Zakká Mathieu Morlon

Google DeepMind

2024-06-27

Gemma 2: Improving Open Language Models at a Practical Size

Gemma Team, Google DeepMind¹

Phi-3 Tech
le Language

Microsoft

1d7db9

Let's look at the data (on dense architectures)

Learn from the many other models (and papers) out there

Model Name	Year	Norm	Parallel Layer	Pre-norm	Position embedding	Activations	Stability tricks
Original transformer	2017	LayerNorm	Serial	□	Sine	ReLU	
GPT	2018	LayerNorm	Serial	□	Absolute	ReLU	
T5 (11B)	2019	RMSNorm	Serial	✓	Relative	ReLU	
GPT2	2019	LayerNorm	Serial	✓	Absolute	ReLU	
T5 (XXL 11B) v1.1	2020	RMSNorm	Serial	✓	Relative	ReLU	
mT5	2020	RMSNorm	Serial	✓	Relative	ReLU	
GPT3 (175B)	2020	LayerNorm	Serial	✓	Absolute	ReLU	
GPTJ	2021	LayerNorm	Parallel	✓	RoPE	ReLU	
LmDQA	2021			✓	Relative	ReLU	
Anthropic LM (not Claude)	2021			✓			
Gopher (280B)	2021	RMSNorm	Serial	✓	Relative	ReLU	
GPT-NeoX	2022	LayerNorm	Parallel	✓	RoPE	ReLU	
BLOOM (175B)	2022	LayerNorm	Serial	✓	Alibi	ReLU	
OPT (175B)	2022	LayerNorm	Serial	✓	Absolute	ReLU	
PoLM (540B)	2022	RMSNorm	Parallel	✓	RoPE	SwiGLU	Z-loss
Chinchilla	2022	RMSNorm	Serial	✓	Relative	ReLU	
Mistral (7B)	2023	RMSNorm	Serial	✓	RoPE	SwiGLU	
LLaMA2 (70B)	2023	RMSNorm	Serial	✓	RoPE	SwiGLU	
LLaMA2 (67B)	2023	RMSNorm	Serial	✓	RoPE	SwiGLU	
GPT4	2023			□			
Olmzo 2	2024	RMSNorm	Serial	□	RoPE	SwiGLU	Z-loss QK-norm
Gemma 2 (27B)	2024	RMSNorm	Serial	✓	RoPE	ReLU	Logit soft capping Pre+post norm
Nemotron-4 (340B)	2024	LayerNorm	Serial	✓	RoPE	SqReLU	
Owen 2 (72B) - same for 2.5	2024	RMSNorm	Serial	✓	RoPE	SwiGLU	
Falcon 2.11B	2024	LayerNorm	Parallel	✓	RoPE	ReLU	Z-loss
Phi3 (small) - same for phi4	2024	RMSNorm	Serial	✓	RoPE	ReLU	
Llama 3 (70B)	2024	RMSNorm	Serial	✓	RoPE	SwiGLU	
Reka Flash	2024	RMSNorm	Serial	✓	RoPE	SwiGLU	
Command R+	2024	LayerNorm	Parallel	✓	RoPE	SwiGLU	
QLMo	2024	RMSNorm	Serial	✓	RoPE	SwiGLU	
Owen (14B)	2024	RMSNorm	Serial	✓	RoPE	SwiGLU	
DeepSeek (67B)	2024	RMSNorm	Serial	✓	RoPE	SwiGLU	
YI (34B)	2024	RMSNorm	Serial	✓	RoPE	SwiGLU	
Mixtial of Experts	2024			□			
Command A	2025	LayerNorm	Parallel	✓	Hybrid (RoPE+NoPE)	SwiGLU	
Gemma 3	2025	RMSNorm	Serial	□	RoPE	ReLU	Pre+post norm QK-norm
SmallM2 (1.7B)	2025	RMSNorm	Serial	✓	RoPE	SwiGLU	

people explore & stick to what works, like evolution.

We will talk through many major architecture and hyperparameter variants.

- What do all these models have in common?
- What parts vary?
- What can we learn from this?

What are we going to cover?

Common architecture variations

- Activations, FFN
- Attention variants
- Position embeddings

Hyperparameters that (do or don't) matter

- What is ff_dim? Do multi_head dims always sum to model_dim?
- How many vocab elements?

Stability tricks

Architecture variations..

Let's think about the core architecture piece

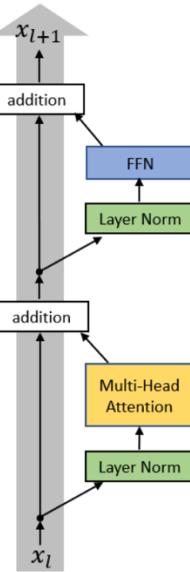
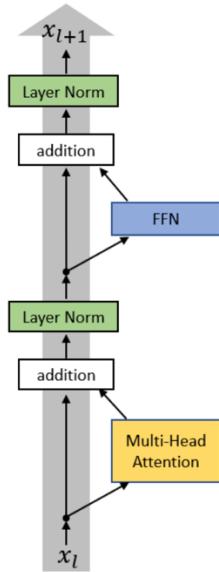
Aa Name	#	Year	Norm	Parallel Layer	Pre-norm	Position embedding	Activations
Original transformer		2017	LayerNorm	Serial	<input type="checkbox"/>	Sine	ReLU
GPT		2018	LayerNorm	Serial	<input type="checkbox"/>	Absolute	GeLU
T5 (11B)		2019	RMSNorm	Serial	<input checked="" type="checkbox"/>	Relative	ReLU
GPT2		2019	LayerNorm	Serial	<input checked="" type="checkbox"/>	Sine	GeLU
T5 (XXL 11B) v1.1		2020	RMSNorm	Serial	<input checked="" type="checkbox"/>	Relative	GeGLU
mT5		2020	RMSNorm	Serial	<input checked="" type="checkbox"/>	Relative	GeGLU
GPT3 (175B)		2020	LayerNorm	Serial	<input checked="" type="checkbox"/>	Sine	GeLU
GPTJ		2021	LayerNorm	Parallel	<input checked="" type="checkbox"/>	RoPE	GeLU
LaMDA		2021			<input checked="" type="checkbox"/>	Relative	GeGLU
Gopher (280B)		2021	RMSNorm	Serial	<input checked="" type="checkbox"/>	Relative	ReLU
GPT-NeoX		2022	LayerNorm	Parallel	<input checked="" type="checkbox"/>	RoPE	GeLU
BLOOM (175B)		2022	LayerNorm	Serial	<input checked="" type="checkbox"/>	Alibi	GeLU
OPT (175B)		2022	LayerNorm	Serial	<input checked="" type="checkbox"/>	Absolute	ReLU
PaLM (540B)		2022	RMSNorm	Parallel	<input checked="" type="checkbox"/>	RoPE	SwiGLU
Chinchilla		2022	RMSNorm	Serial	<input checked="" type="checkbox"/>	Relative	ReLU
Mistral (7B)		2023	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
LLaMA2 (70B)		2023	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
LLaMA (65B)		2023	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
Qwen (14B)		2024	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
DeepSeek (67B)		2024	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
Yi (34B)		2024	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU

High level view:

- Low consensus
(except pre-norm)
- Trends toward 'LLaMA-like' architectures

Pre-vs-post norm

The one thing everyone agrees on (in 2024)



2017 paper

Post-LN Transformer

$$\begin{aligned}x_{l,i}^{post,1} &= \text{MultiHeadAtt}(x_{l,i}^{post}, [x_{l,1}^{post}, \dots, x_{l,n}^{post}]) \\x_{l,i}^{post,2} &= x_{l,i}^{post} + x_{l,i}^{post,1} \\x_{l,i}^{post,3} &= \text{LayerNorm}(x_{l,i}^{post,2}) \\x_{l,i}^{post,4} &= \text{ReLU}(x_{l,i}^{post,3}W^{1,l} + b^{1,l})W^{2,l} + b^{2,l} \\x_{l,i}^{post,5} &= x_{l,i}^{post,3} + x_{l,i}^{post,4} \\x_{l+1,i}^{post} &= \text{LayerNorm}(x_{l,i}^{post,5})\end{aligned}$$

Pre-LN Transformer

$$\begin{aligned}x_{l,i}^{pre,1} &= \text{LayerNorm}(x_{l,i}^{pre}) \\x_{l,i}^{pre,2} &= \text{MultiHeadAtt}(x_{l,i}^{pre,1}, [x_{l,1}^{pre,1}, \dots, x_{l,n}^{pre,1}]) \\x_{l,i}^{pre,3} &= x_{l,i}^{pre} + x_{l,i}^{pre,2} \\x_{l,i}^{pre,4} &= \text{LayerNorm}(x_{l,i}^{pre,3}) \\x_{l,i}^{pre,5} &= \text{ReLU}(x_{l,i}^{pre,4}W^{1,l} + b^{1,l})W^{2,l} + b^{2,l} \\x_{l+1,i}^{pre} &= x_{l,i}^{pre,5} + x_{l,i}^{pre,3}\end{aligned}$$

Final LayerNorm: $x_{Final,i}^{pre} \leftarrow \text{LayerNorm}(x_{L+1,i}^{pre})$

all recent ones

Set up LayerNorm so that it doesn't affect the main residual signal path (on the left)

Figure from Xiong 2020

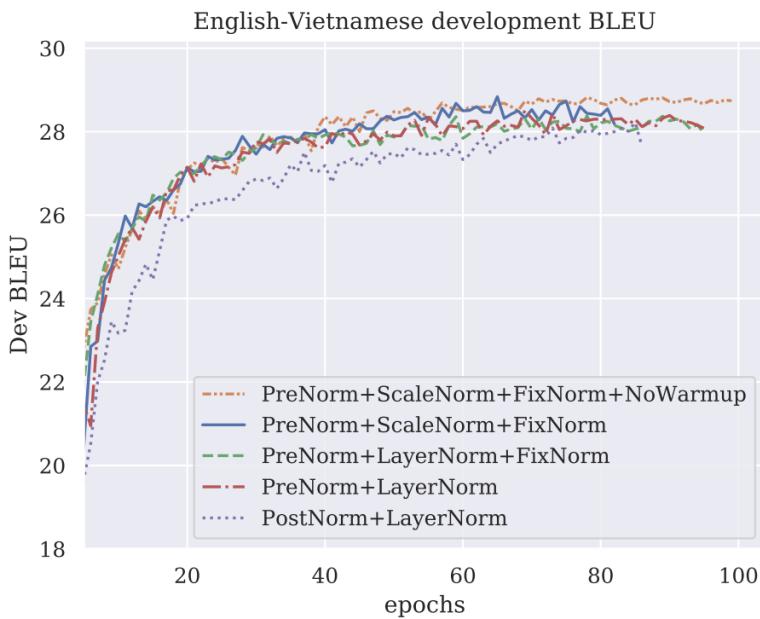
Almost all modern LMs use pre-norm (but BERT was post-norm)

(One somewhat funny exception – OPT350M. I don't know why this is post-norm)

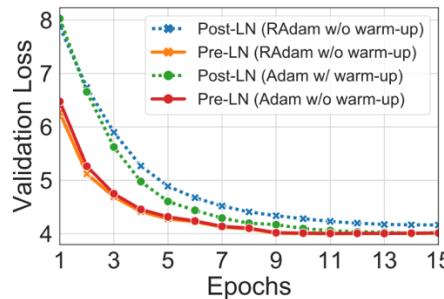
original
trans form
post-norm

Pre-vs-post-norm, the data

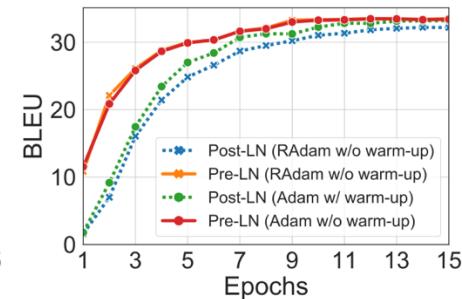
post-norm needs tricks to work & is not super stable.



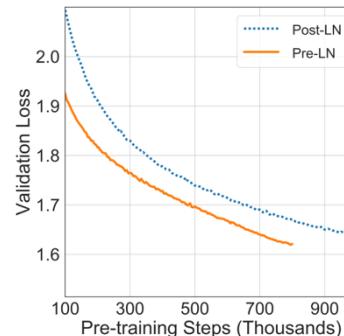
Salazar and Ngyuen 2019



(a) Validation Loss (IWSLT)



(b) BLEU (IWSLT)



(a) Validation Loss on BERT

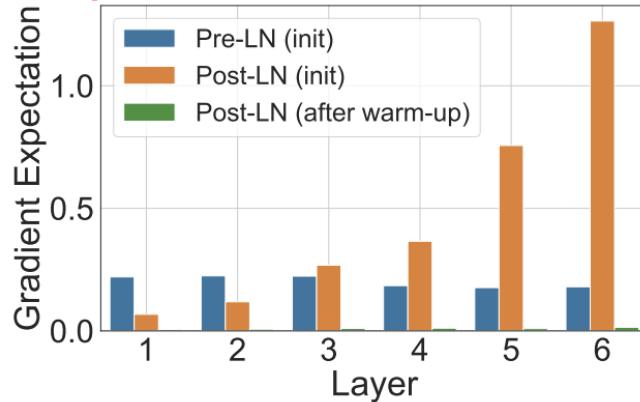
Figure from Xiong 2020

Pre-vs-post norm, explanations?

pre-norm \rightarrow gradients remain balanced.

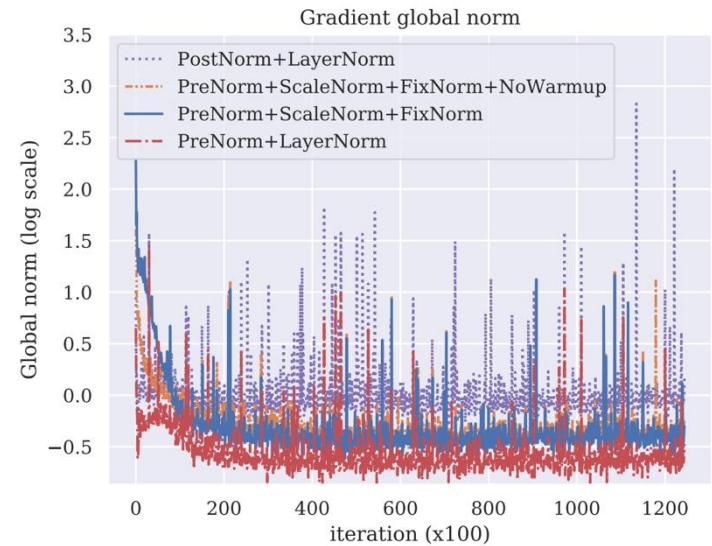
Gradient attenuation [Xiong 2020]

post-norm \rightarrow gradients shoot up!



(a) W^1 in the FFN sub-layers

Gradient spikes [Salazar and Ngyuen]

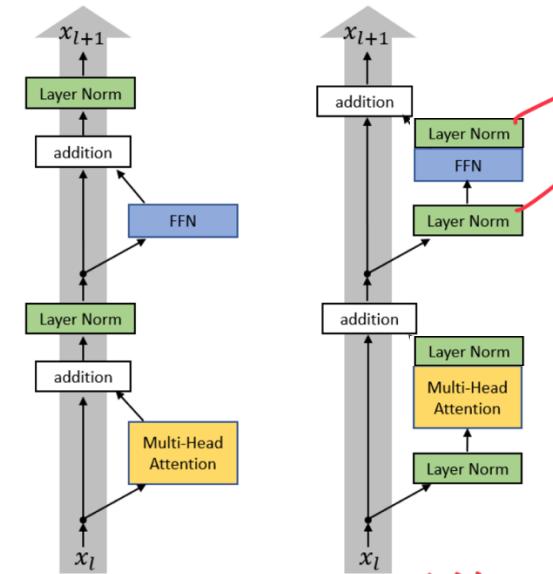


Original stated advantage – removing warmup.
Today – stability and larger LRs for large networks

New things – ‘double’ norm.

If putting LayerNorms in residual streams is bad.. Why not post-norm outside the stream?

gradient flowing backward
has to interact with
these blocks. so, large
models might not
train well like LSTM,
SSMs



why front, why not
at back!!

→ does it both front & back!

Recent models: Grok, Gemma 2. Olmo 2 only does non-residual post norm

LayerNorm vs RMSNorm

Original transformer: **LayerNorm** – normalizes the mean and variance across d_{model}

Notable models:

GPT3/2/1, OPT, GPT-J, BLOOM

$$y = \frac{x - \mathbb{E}[x]}{\sqrt{\text{Var}[x] + \epsilon}} * \gamma + \beta$$

normalize the activations.

Many modern LMs: **RMSNorm** – does not subtract mean or add a bias term

Notable models:

LLaMA-family, PaLM, Chinchilla, T5

$$y = \frac{x}{\sqrt{\|x\|_2^2 + \epsilon}} * \gamma$$

not shifting activations
is generally better

Why RMSNorm?

Modern explanation – it's faster (and just as good).

- **Fewer operations** (no mean calculation), *no bias addition*.
- **Fewer parameters** (no bias term to store)

$$y = \frac{x - \mathbb{E}[x]}{\sqrt{\text{Var}[x] + \epsilon}} * \gamma + \beta$$

Does this explanation make sense?

Operator class	% flop
△ Tensor contraction	99.80
□ Stat. normalization	0.17
○ Element-wise	0.03

→ matrix mult.

Matrix multiplies are the *vast* majority of FLOPs (and memory)

Memory move
movement is all
we need.
[Ivanov et al 2023]

Why RMSNorm (2)

FLOPs are for multiplications.
data movement also takes time.

Important lesson: FLOPS are not runtime! (we will discuss this in far more detail later)

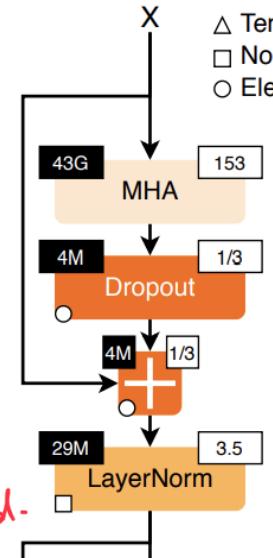
Operator class	% flop	% Runtime
△ Tensor contraction	99.80	61.0 -
□ Stat. normalization	0.17	25.5
○ Element-wise	0.03	13.5 - these

These 0.17% of operations take
25.5% of runtime bcoz of
memory bound overhead.

RMSNorm can still matter due to
the importance of *data movement*



movement
memory is
also imp. comp. to
FLOPs



Left top ("43G") is FLOPS

Right top ("153") is the FLOP-to-memory ratio

RMSNorm - validation

RMSNorm runtime (and surprisingly, perf) gains have been seen in papers

Model	Params	Ops	Step/s	Early loss	Final loss	SGLUE	XSum	WebQ	WMT	EnDe
Vanilla Transformer	$223M$	$11.1T$	3.50	2.182 ± 0.005	1.838	71.66	17.78	23.02		26.62
RMS Norm	$223M$	$11.1T$	3.68	2.167 ± 0.008	1.821	75.45	17.94	24.07		27.14
Rezero	$223M$	$11.1T$	3.51	2.262 ± 0.003	1.939	61.69	15.64	20.90		26.37
Rezero + LayerNorm	$223M$	$11.1T$	3.26	2.223 ± 0.006	1.858	70.42	17.58	23.02		26.29
Rezero + RMS Norm	$223M$	$11.1T$	3.34	2.221 ± 0.009	1.875	70.33	17.32	23.02		26.19
Fixup	$223M$	$11.1T$	2.95	2.382 ± 0.012	2.067	58.56	14.42	23.02		26.31

Narang et al 2020

More generally: dropping bias terms

Most modern transformers don't have bias terms.

as it's additional
memory movement
time & can be
learn without it.

operations (↓)

but FLOPs
are very less
already comp.
to mat muls)

Original Transformer:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Most implementations (if they're not gated):

$$\text{FFN}(x) = \sigma(xW_1)W_2$$

Reasons: memory (similar to RMSnorm) and optimization stability

LayerNorm: recap

- Basically everyone does pre-norm.
 - Intuition – keep the good parts of residual connections
 - Observations – nicer gradient propagation, fewer spike
 - Some people add a second norm outside the residual stream (NOT post-norm)
 - Most people do RMSnorm
 - In practice, works as well as LayerNorm
 - But, has fewer parameters to move around, which saves on wallclock time
 - People more generally drop bias terms since the compute/param tradeoffs are not great.
- put layer norm (or RMSnorm)
outside the residual
line.
↑
Pre-norm helps
gradients to be
in check.

Activations

A whole zoo of activations ..

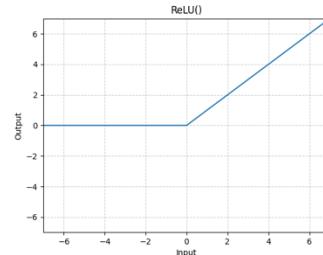
ReLU, GeLU, Swish, ELU, GLU, GeGLU, ReGLU, SeLU, SwiGLU, LiGLU

What are these things? What do people use? Does it matter?

A few of the common activations

ReLU

$$FF(x) = \max(0, xW_1) W_2$$



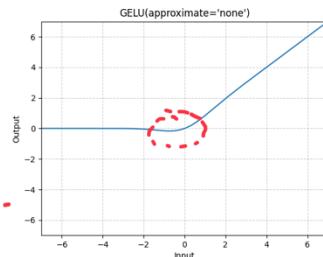
gaussian error
linear unit

GeLU

$$FF(x) = \text{GELU}(xW_1)W_2$$

$$\text{GELU}(x) := x\Phi(x)$$

makes more
differentiable



SwiGLU / GeGLU (next slide..)

Notable models:

Original transformer, T5,
Gopher, Chinchilla, OPT

Notable models:

GPT1/2/3, GPTJ, GPT-Neox,
BLOOM

Notable models:

Llama, PaLM, T5 v1.1, most
models post 2023

Gated activations (*GLU)

(gating)

GLUs modify the ‘first part’ of a FF layer

$$FF(x) = \max(0, xW_1) W_2$$

basically gate the op with

Instead of a linear + ReLU, augment the above with an (entrywise) linear term

$$\max(0, xW_1) \rightarrow \max(0, xW_1) \otimes (xV)$$

GLU + ReLU ↳

This gives the gated variant (ReGLU) – note that we have an extra parameter (V)

$$FF_{\text{ReGLU}}(x) = \underbrace{\max(0, xW_1) \otimes xV}_{\text{gated the hidden part of mlp}} W_2$$

Gated variants of standard FF layers

GeGLU

$$\text{FFN}_{\text{GeGLU}}(x, W, V, W_2) = (\text{GELU}(xW) \otimes \underline{xV})W_2$$

optimized implementations might make
these faster

SwiGLU (swish is $x * \text{sigmoid}(x)$)

(FLOPs-wiser, it's negligible anyway)

$$\text{FFN}_{\text{SwiGLU}}(x, W, V, W_2) = (\text{Swish}_1(xW) \otimes xV)W_2$$

$$((\overset{\downarrow}{xW}) \text{sig}(\overset{\downarrow}{xW}) \otimes \overset{\downarrow}{xV})W_2$$

Note: Gated models use smaller dimensions for the d_{ff} by 2/3

compared to non gated counterparts

so that this extra 'V' params still make total params the same

Notable models:

T5 v1.1, mT5, LaMDA, Phi3,
Gemma 2, Gemma 3

Notable models:

LLaMa 1/2/3, PaLM, Mistral,
OIMo, most models post 2023

Do gated linear units work?

Yes, fairly consistently so.

	Score	CoLA	SST-2
	Average	MCC	Acc
FFN _{ReLU}	83.80	51.32	94.04
FFN _{GELU}	83.86	53.48	94.04
FFN _{Swish}	83.60	49.79	93.69
FFN _{GLU}	84.20	49.16	94.27
FFN _{GEGLU}	84.12	53.65	93.92
FFN _{Bilinear}	83.79	51.02	94.38
FFN _{SwiGLU}	84.36	51.59	93.92
FFN _{ReGLU}	84.67	56.16	94.38
[Raffel et al., 2019]	83.28	53.84	92.68
ibid. stddev.	0.235	1.111	0.569

Shazeer 2020

Do gated linear units work (2)?

Yes, with other works corroborating Shazeer 2020

Model	Params	Ops	Step/s	Early loss	Final loss	SGLUE	XSum	WebQ
Vanilla Transformer	$223M$	$11.1T$	3.50	2.182 ± 0.005	1.838	71.66	17.78	23.02
GeLU	$223M$	$11.1T$	3.58	2.179 ± 0.003	1.838	75.79	17.86	25.13
Swish	$223M$	$11.1T$	3.62	2.186 ± 0.003	1.847	73.77	17.74	24.34
ELU	$223M$	$11.1T$	3.56	2.270 ± 0.007	1.932	67.83	16.73	23.02
GLU	$223M$	$11.1T$	3.59	2.174 ± 0.003	1.814	74.20	17.42	24.34
GeGLU	$223M$	$11.1T$	3.55	2.130 ± 0.006	1.792	75.96	18.27	24.87
ReGLU	$223M$	$11.1T$	3.57	2.145 ± 0.004	1.803	76.17	18.36	24.87
SeLU	$223M$	$11.1T$	3.55	2.315 ± 0.004	1.948	68.76	16.76	22.75
SwiGLU	$223M$	$11.1T$	3.53	2.127 ± 0.003	1.789	76.00	18.20	24.34
LiGLU	$223M$	$11.1T$	3.59	2.149 ± 0.005	1.798	75.34	17.97	24.34
Sigmoid	$223M$	$11.1T$	3.63	2.291 ± 0.019	1.867	74.31	17.51	23.02
Softplus	$223M$	$11.1T$	3.47	2.207 ± 0.011	1.850	72.45	17.65	24.34

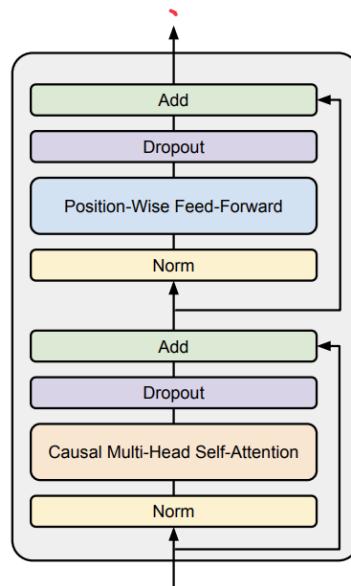
Narang et al 2020

Gating, activations

- **Many variations (ReLU, GeLU, *GLU) across models.**
- ***GLU isn't necessary for a good model (see GPT3), but it's probably helpful**
Other, recent outlier models..
Nemotron 340B (Squared ReLU), Falcon 2 11b (ReLU)
- **But evidence points towards somewhat consistent gains from Swi/GeGLU**

Serial vs Parallel layers

Normal transformer blocks are *serial* – they compute attention, then the MLP



Serial connections
↓
lower GPU utilization.

$$y = \text{norm}(\text{FF}(\text{norm}(x + \text{At}(\text{norm}(x))))))$$

Could we parallelize the transformer block?

Parallel layers

A few models (GPTJ, PaLM, GPT-NeoX) do parallel layers. Originally in GPT-J

Parallel Layers – We use a “parallel” formulation in each Transformer block ([Wang & Komatsuzaki, 2021](#)), rather than the standard “serialized” formulation. Specifically, the standard formulation can be written as:

$$y = x + \text{MLP}(\text{LayerNorm}(x + \text{Attention}(\text{LayerNorm}(x))))$$

Whereas the parallel formulation can be written as:

$$y = x + \text{MLP}(\text{LayerNorm}(x)) + \text{Attention}(\text{LayerNorm}(x))$$

The parallel formulation results in roughly 15% faster training speed at large scales, since the MLP and Attention input matrix multiplications can be fused. Ablation experiments showed a small quality degradation at 8B scale but no quality degradation at 62B scale, so we extrapolated that the effect of parallel layers should be quality neutral at the 540B scale.

If implemented right, LayerNorm can be shared, and matrix multiplies can be fused

Recent Models: Cohere Command A, Falcon 2 11B, Command R+

Summary: architectures

Pre-vs-post norm:

- Everyone does pre-norm (except OPT350M), likely with good reason.

Layer vs RMSnorm:

- RMSnorm has clear compute wins, sometimes even performance

Gating:

- GLUs seem generally better, though differences are small

Serial vs parallel layers:

- No extremely serious ablations, but has a compute win.

All Name	# Year	Norm	Parallel Layer	Pre-norm	Position embedding	Activations
Original transformer	2017	LayerNorm	Serial	<input type="checkbox"/>	Sine	ReLU
GPT	2018	LayerNorm	Serial	<input type="checkbox"/>	Absolute	GeLU
T5 (11B)	2019	RMSNorm	Serial	<input checked="" type="checkbox"/>	Relative	ReLU
GPT2	2019	LayerNorm	Serial	<input checked="" type="checkbox"/>	Absolute	GeLU
T5 (XXL 11B) v1.1	2020	RMSNorm	Serial	<input checked="" type="checkbox"/>	Relative	GeGLU
mT5	2020	RMSNorm	Serial	<input checked="" type="checkbox"/>	Relative	GeGLU
GPT3 (175B)	2020	LayerNorm	Serial	<input checked="" type="checkbox"/>	Absolute	GeLU
GPTJ	2021	LayerNorm	Parallel	<input checked="" type="checkbox"/>	RoPE	GeLU
LaMDA	2021			<input checked="" type="checkbox"/>	Relative	GeGLU
Anthropic LM (not Claude)	2021			<input checked="" type="checkbox"/>		
Gopher (280B)	2021	RMSNorm	Serial	<input checked="" type="checkbox"/>	Relative	ReLU
GPT-NeoX	2022	LayerNorm	Parallel	<input checked="" type="checkbox"/>	RoPE	GeLU
BLOOM (175B)	2022	LayerNorm	Serial	<input checked="" type="checkbox"/>	AllBi	ReLU
GPT (175B)	2022	LayerNorm	Serial	<input checked="" type="checkbox"/>	Absolute	ReLU
PaLM (640B)	2022	RMSNorm	Parallel	<input checked="" type="checkbox"/>	RoPE	SwiGLU
Chinchilla	2022	RMSNorm	Serial	<input checked="" type="checkbox"/>	Relative	ReLU
Mistral (7B)	2023	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
LLaMA2 (70B)	2023	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
LLaMA (65B)	2023	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
GPT4	2023			<input type="checkbox"/>		
Baichuan 2	2023	RMSNorm	Serial	<input checked="" type="checkbox"/>	AllBi	SwiGLU
Olmo 2	2024	RMSNorm	Serial	<input type="checkbox"/>	RoPE	SwiGLU
Gemma 2 (27B)	2024	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	GeGLU
Nemtron-4 (340B)	2024	LayerNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SqReLU
Qwen 2 (72B) - same for 2.5	2024	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
Falcon 2.11B	2024	LayerNorm	Parallel	<input checked="" type="checkbox"/>	RoPE	GeLU
Phi3 (small) - same for phi4	2024	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	GeGLU
Llama 3 (70B)	2024	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
Reka Flash	2024	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
Command R+	2024	LayerNorm	Parallel	<input checked="" type="checkbox"/>	RoPE	SwiGLU
OLMo	2024	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
Qwen (14B)	2024	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
DeepSeek (67B)	2024	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
YI (34B)	2024	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU
Mixtral of Experts	2024			<input type="checkbox"/>		
Command A	2025	LayerNorm	Parallel	<input checked="" type="checkbox"/>	Hybrid (RoPE+NoPE)	SwiGLU
Gemma 3	2025	RMSNorm	Serial	<input type="checkbox"/>	RoPE	GeGLU
SmallLM2 (1.7B)	2025	RMSNorm	Serial	<input checked="" type="checkbox"/>	RoPE	SwiGLU

Many variations in position embeddings

Sine embeddings: add sines and cosines that enable localization

$$Embed(x, i) = v_x + PE_{pos}$$

$$PE_{(pos, 2i)} = \sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos, 2i+1)} = \cos(pos/10000^{2i/d_{\text{model}}})$$

Notable models:

Original transformer

Absolute embeddings: add a position vector to the embedding

$$Embed(x, i) = v_x + u_i$$

Notable models:

GPT1/2/3, OPT

Relative embeddings: add a vector to the *attention computation*

$$e_{ij} = \frac{x_i W^Q (x_j W^K + a_{ij}^K)^T}{\sqrt{d_z}}$$

Notable models:

T5, Gopher, Chinchilla

Rope embeddings (next slides..)

Notable models:

GPTJ, PaLM, LLaMA
Most 2024+ models

Relative
position
embeddings

RoPE: rotary position embeddings

$\alpha \rightarrow \text{word}$, $i \rightarrow \text{position}$
 $f \rightarrow \text{embedding}$.

High level thought process: a relative position embedding should be some $f(x, i)$ s.t.

definition that
enforces absolute
position invariance

pay att. to how
far 2 words are
rather than where ^{they} are.

That is, the attention function only gets to depend on the relative position ($i-j$). How do existing embeddings not fulfill this goal?

- **Sine:** Has various cross-terms that are not relative

$$\langle \text{Embed}(x, i), \text{Embed}(y, i) \rangle = \langle v_x, v_y \rangle + \langle PE_i, v_y \rangle \dots$$

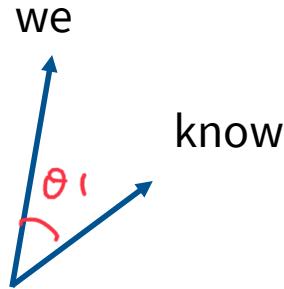
- **Absolute:** obviously not relative

- **Relative embeddings:** $e_{ij} = \frac{x_i W^Q (x_j W^K + a_{ij}^K)^T}{\sqrt{d_z}}$ is not an inner product

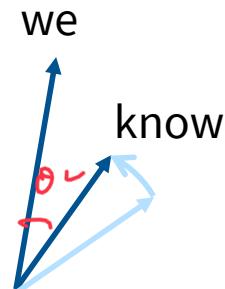
RoPE: rotary position embeddings

How can we solve this problem?

- We want our embeddings to be invariant to absolute position
- We know that inner products are invariant to arbitrary rotation.

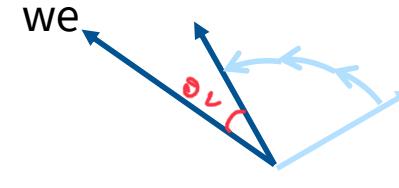


Position independent embedding



Embedding "we know that"
Rotate we by '0 positions'
know by '1 positions'

angle is preserved
as we know distance
stays same!



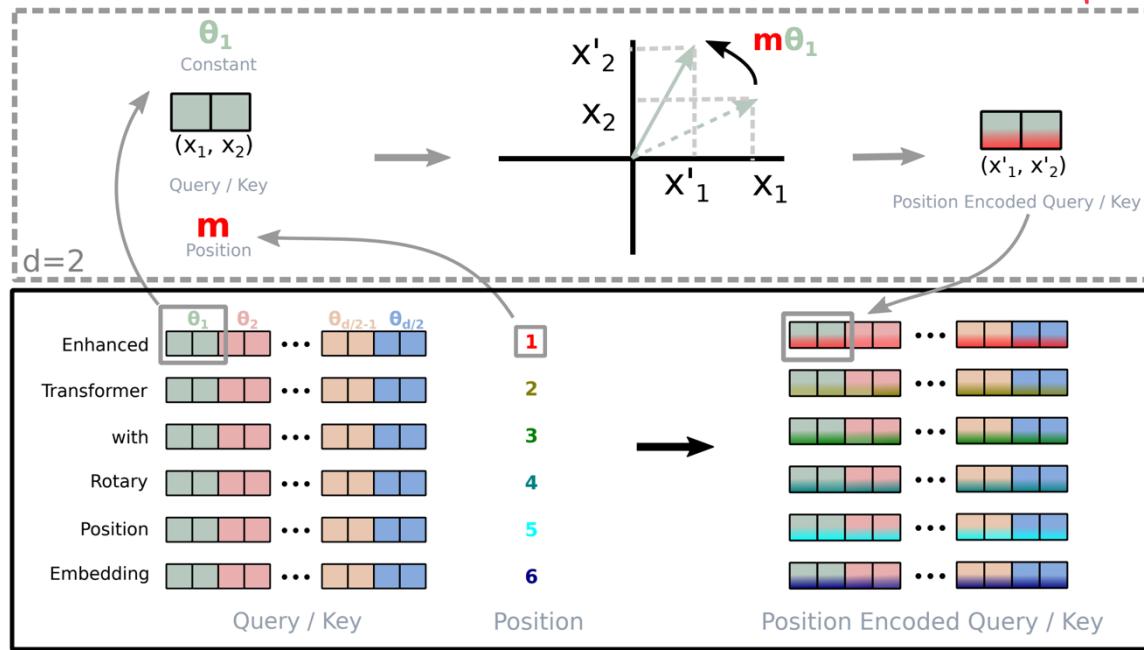
Embedding "of course we know"
Rotate we by '2 positions'
know by '3 positions'

dif ①

RoPE: rotary position embeddings

There are many rotations, which one do you pick?

bcoz in 2D it's clear
but in nD there're
 n options!



'D' is cut into blocks of 2-dim. & each one is rotated. they are rotated diff ($\Theta_1, \Theta_2, \dots$) so that some are rotated quickly while others are slowly [Su et al 2021]

Just pair up the coordinates and rotate them in 2d (motivation: complex numbers)

The actual RoPE math

Multiply with sines and cosines \rightarrow rotation in embedding space.

these θ 's are not hypers but are
a schedule (lookup)

$$f_{\{q,k\}}(\mathbf{x}_m, m) = \mathbf{R}_{\Theta,m}^d \mathbf{W}_{\{q,k\}} \mathbf{x}_m \quad (14)$$

table
selected by
authors

it's just one
additional
matmul.

$$\mathbf{R}_{\Theta,m}^d = \begin{pmatrix} \cos m\theta_1 & -\sin m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ \sin m\theta_1 & \cos m\theta_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cos m\theta_2 & -\sin m\theta_2 & \cdots & 0 & 0 \\ 0 & 0 & \sin m\theta_2 & \cos m\theta_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & \cos m\theta_{d/2} & -\sin m\theta_{d/2} \\ 0 & 0 & 0 & 0 & \cdots & \sin m\theta_{d/2} & \cos m\theta_{d/2} \end{pmatrix} \quad (15)$$

as we have divided
 d to 2 dim. blocks.

Difference with sine embeddings – not additive, no cross terms

Implementation and code for RoPE

Usual
attention stuff

```
query_states = self.q_proj(hidden_states)
key_states = self.k_proj(hidden_states)
value_states = self.v_proj(hidden_states)

# Flash attention requires the input to have the shape
# batch_size x seq_length x head_dim x hidden_dim
# therefore we just need to keep the original shape
query_states = query_states.view(bsz, q_len, self.num_heads, self.head_dim).transpose(1, 2)
key_states = key_states.view(bsz, q_len, self.num_key_value_heads, self.head_dim).transpose(1, 2)
value_states = value_states.view(bsz, q_len, self.num_key_value_heads, self.head_dim).transpose(1, 2)
```

Get the RoPE
matrix cos/sin

```
cos, sin = self.rotary_emb(value_states, position_ids)
query_states, key_states = apply_rotary_pos_emb(query_states, key_states, cos, sin)
```

Multiply
query/key inputs

...

Same stuff as the usual multi-head self attention below

we don't do RoPE at bottom like pos. emb. but at every attention block.

Note: embedding at each attention operation to enforce position invariance

where a token is there is not imp. compared to relative to others!

Hyperparameters

Transformer hyperparameter questions you might have had in 224n..

- How much bigger should the feedforward size be compared to hidden size?
- How many heads, and should num_heads always divide hidden size?
- What should my vocab size be?

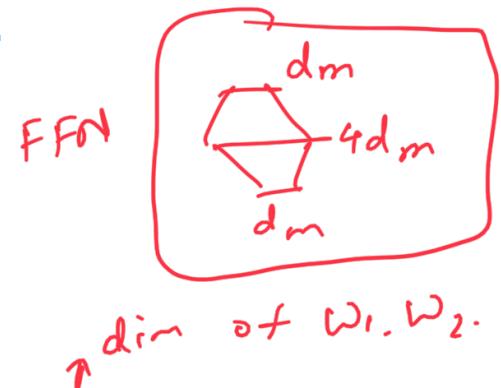
And other model setting questions

- Do people even regularize these huge LMs?
- How do people scale these models - very deep or very wide?

Surprising (?) consensus hyperparameter 1

Feedforward – model dimension ratio.

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$



There are two dimensions that are relevant – the feedforward dim (d_{ff}) and model dim (d_{model}). What should their relationship be?

dim. of $x.$

(batch size $\propto d_{model}$)

$$d_{ff} = 4 d_{model}$$

$$(d_{model} \times d_{ff})$$

This is *almost always* true. There's just a few exceptions.

Exception #1 – GLU variants *to make parameters count same across models.*

Remember that GLU variants scale down by 2/3rd. This means most GLU variants have

$$d_{ff} = \frac{8}{3} d_{model}$$
. This is mostly what happens. Some notable such examples.

Model	d_{ff}/d_{model}
PaLM	4
Mistral 7B	3.5
LLaMA-2 70B	3.5
LLaMA 70B	2.68
Qwen 14B	2.67
DeepSeek 67B	2.68
Yi 34B	2.85
T5 v1.1	2.5

$$\left. \right\} \frac{8}{3}$$

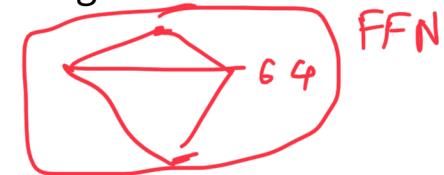
Models are roughly in this range, though PaLM, LLaMA2 and Mistral are slightly larger

Exception #2 – T5

As we have (and will) see, most LMs are have boring, conservative hyperparameters. One exception is T5 [Raffel et al 2020] which has some *very bold* settings.

In particular, for the 11B model, they set

$$\begin{aligned} d_{ff} &= 65,536 \\ d_{model} &= 1024 \end{aligned} \quad \left. \right\} 64 \times$$



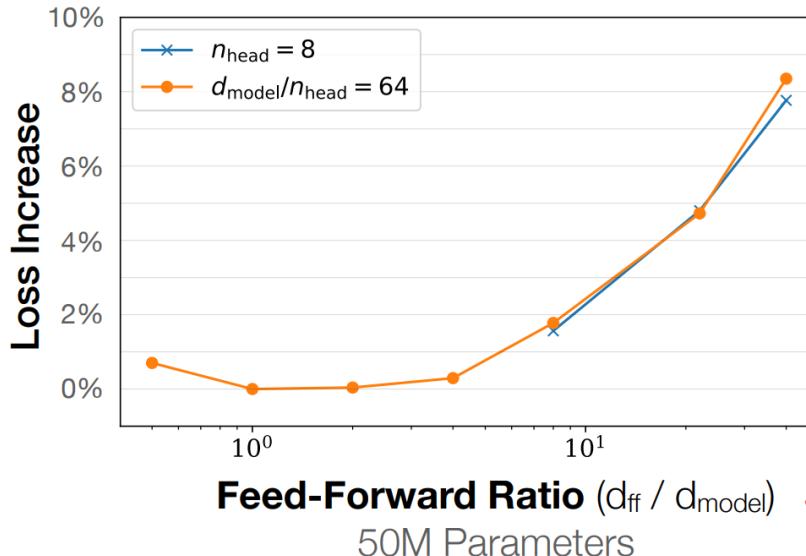
For an astounding 64-times multiplier.

for “11B” we use $d_{ff} = 65,536$ with 128-headed attention producing a model with about 11 billion parameters. We chose to scale up d_{ff} specifically because modern accelerators (such as the TPUs we train our models on) are most efficient for large dense matrix multiplications like those in the Transformer’s feed-forward networks.

Other, recent exceptions – Gemma 2 (8x), SmoLLM/Gemma 3 (4x, GLU)

Why this range of multipliers?

Empirically, there's a basin between 1-10 where this hyperparameter is near-optimal



scaling law for
hyperparameters.
 $\frac{d_{\text{ff}}}{d_{\text{model}}} \quad \& \quad \frac{d_{\text{model}}}{n_{\text{head}}}$

$\rightarrow 1 - 10$ seems good.

What can we learn from the model-dim hyperparam?

④ wider matrices \rightarrow systems gain of red Comp. but suboptimal in terms of expressivity.

- The ‘default’ choices of $d_{ff} = 4d_{model}$ and $d_{ff} = 2.66d_{model}$ have worked well for nearly all modern LLMs.

⑤ thin matrix but multiple of them \rightarrow no systems gain as it's serial but expressive.

- But T5 does show that even radical choices of $d_{ff} = 64d_{model}$ can work. This hyperparameter choice isn’t written in stone.
- That said, T5 has a follow-up model (T5 v1.1) that is ‘improved’ and uses a much more standard 2.5 multiplier on GeGLU, so the 64-times multiplier is likely suboptimal.



For same parameter count, design choices we can have.

Surprising (?) consensus hyperparameter 2

Head-dim*num-heads to model-dim ratio. As a reminder, slide from 224n.

Multi-head self-attention is computationally efficient

- Even though we compute h many attention heads, it's not really more costly.
 - We compute $XQ \in \mathbb{R}^{n \times d}$, and then reshape to $\mathbb{R}^{n \times h \times d/h}$. (Likewise for XK, XV .)
 - Then we transpose to $\mathbb{R}^{h \times n \times d/h}$; now the head axis is like a batch axis.
 - Almost everything else is identical, and the **matrices are the same sizes**.

This doesn't have to be true: we can have head-dimensions > model-dim / num-heads.

But most models do follow this guideline

we don't have
+ break head-
dim, we can let
it grow

How many heads, whats the model dim?

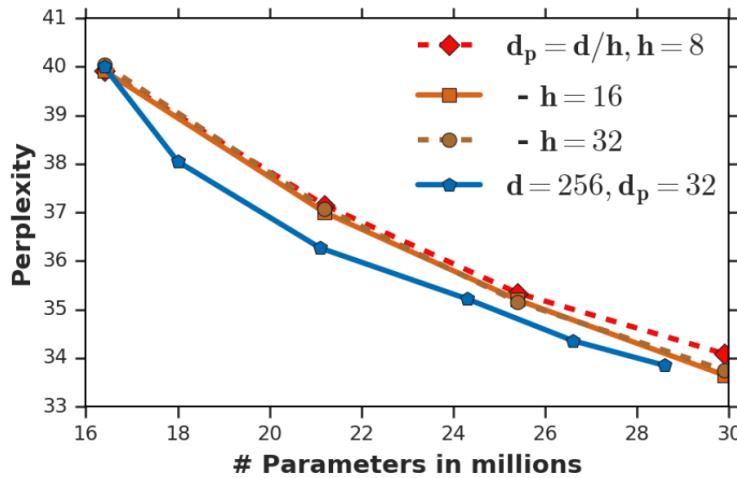
Some examples of this hyperparameter

	Num heads	Head dim	Model dim	Ratio
GPT3	96	128	12288	1
T5	128	128	1024	16
T5 v1.1	64	64	4096	1
LaMDA	128	128	8192	2
PaLM	48	258	18432	1.48
LLaMA2	64	128	8192	1

Most models have ratios around 1 – notable exceptions by some google models.

Evidence for 1-1 ratio?

There have been papers written against the 1-1 ratio [Bhojanapalli et al 2020]



more & more heads \rightarrow low ranks
& less dim for heads effects the expressivity!

theory & practice might not agree!

But we don't seem to be seeing significant 'low rank bottlenecks' in practice..

Aspect ratios

Should my model be deep or wide? *How deep and how wide?*

Most models are surprisingly consistent on this one too!

Model	d_{model}/n_{layer}
BLOOM	205
T5 v1.1	171
PaLM (540B)	156
GPT3/OPT/Mistral/Qwen	128
LLaMA / LLaMA2 / Chinchila	102
T5 (11B)	43
GPT2	33

Sweet spot?

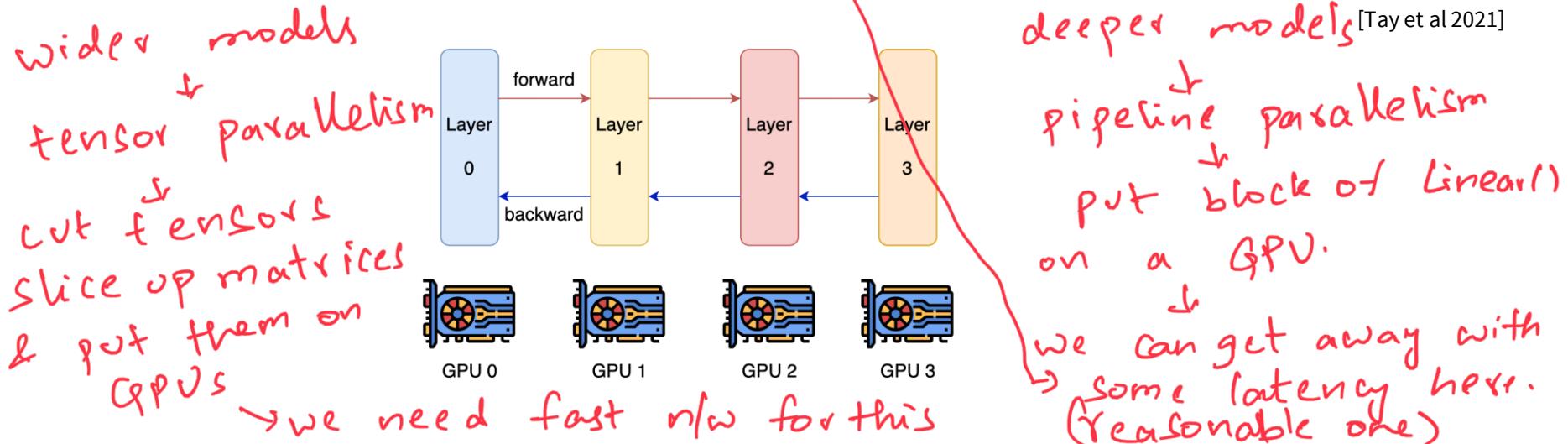
Handwritten notes:
d_{model} (↑) wider.
n_{layer} (↑) deeper.
→ much wider than
deep ↴

} much deeper
than wider.

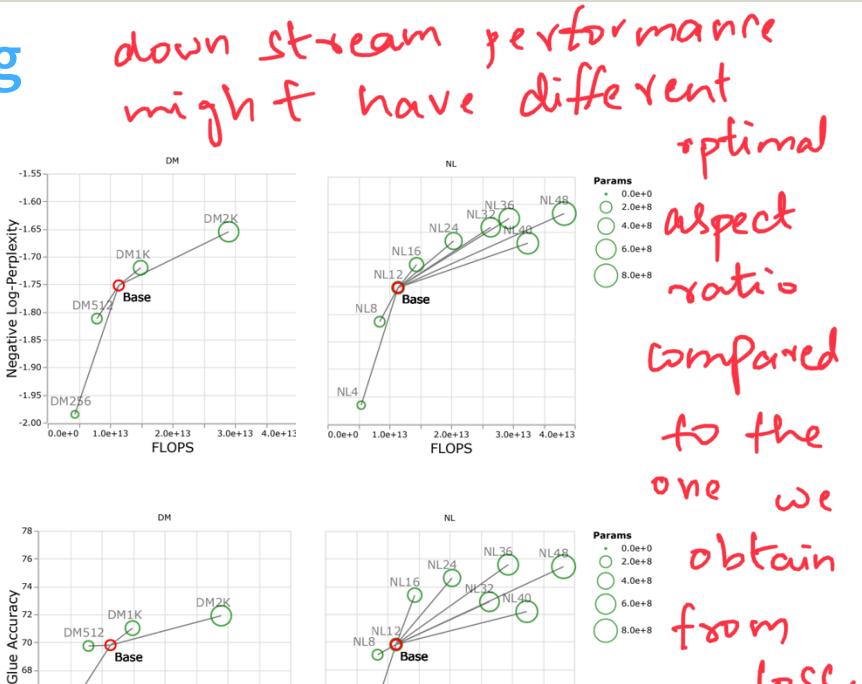
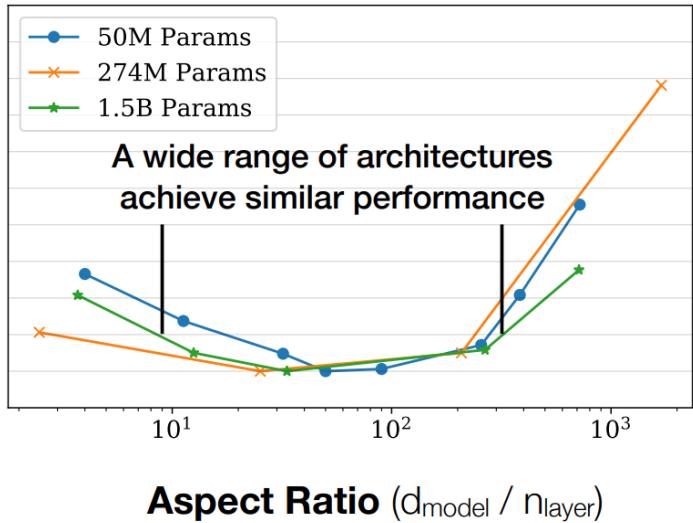
Considerations about aspect ratio

Extremely deep models are harder to parallelize and have higher latency

The Limits of Depth vs Width We note an obvious limitation with our advice. Scaling depth has an obvious limiter, i.e., they are non-parallelizable across different machines or devices and every computation has to always wait for the previous layer. This is unlike width, which can be easily parallelizable over thousands or hundreds of thousands of devices. Within the limitation of scaling



Evidence on aspect ratio scaling



[Kaplan et al 2020]

[Tay et al 2021]

What are typical vocabulary sizes?

Monolingual models – 30-50k vocab

Model	Token count
Original transformer	37000
GPT	40257
GPT2/3	50257
T5/T5v1.1	32128
LLaMA	32000

Common
lot of info packed into
lesser tokens → less inference
costs

Monolingual vocab don't need to be huge, but multilingual ones do

as people interact
expect diff. things.
Multilingual / production systems 100-250k

Model	Token count
mT5	250000
PaLM	256000
GPT4	100276
Command A	255000
DeepSeek	100000
Qwen 15B	152064
Yi	64000

Dropout and other regularization

Do we need regularization during pretraining?

bcoz we don't even overfit,
is that really needed??

Arguments against:

- There is *a lot* of data (trillions of tokens), more than parameters.
- SGD only does a single pass on a corpus (hard to memorize)

This is all quite reasonable.. but what do people do in practice?

Dropout and weight decay in practice

Model	Dropout*	Weight decay
Original transformer	0.1	0
GPT2	0.1	0.1
T5	0.1	0
GPT3	0.1	0.1
T5 v1.1	0	0
PaLM	0	(variable)
OPT	0.1	0.1
LLaMA	0	0.1
Qwen 14B	0.1	0.1

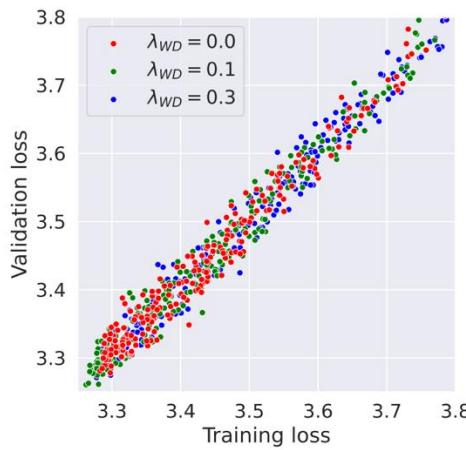
Many older models used dropout during pretraining

Newer models (except Qwen) rely only on weight decay

* Most of the times papers just don't discuss dropout. On open models, this closely matches not doing dropout. This may not be true of closed models.

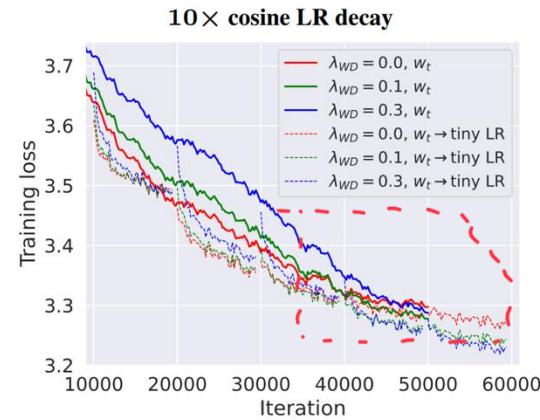
Why weight decay LLMs?

[Andriushchenko et al 2023] has interesting observations about LLM weight decay



It's not to control overfitting

bcoz we do only 1 epoch
train to valid loss stays consistent



Weight decay interacts with learning rates (cosine schedule)

(in an unintuitive way)
at the end of training,
W & WD interact to get
better losses!!

Summary: hyperparameters

Feedforward

- Factor-of-4 rule of thumb (8/3 for GLUs) is standard (with some evidence)

Head dim

- Head dim * Num head = D model is standard
 - but low to no validation

Aspect ratio

- Wide range of ‘good’ values (100-200).
Systems concerns dictate the value

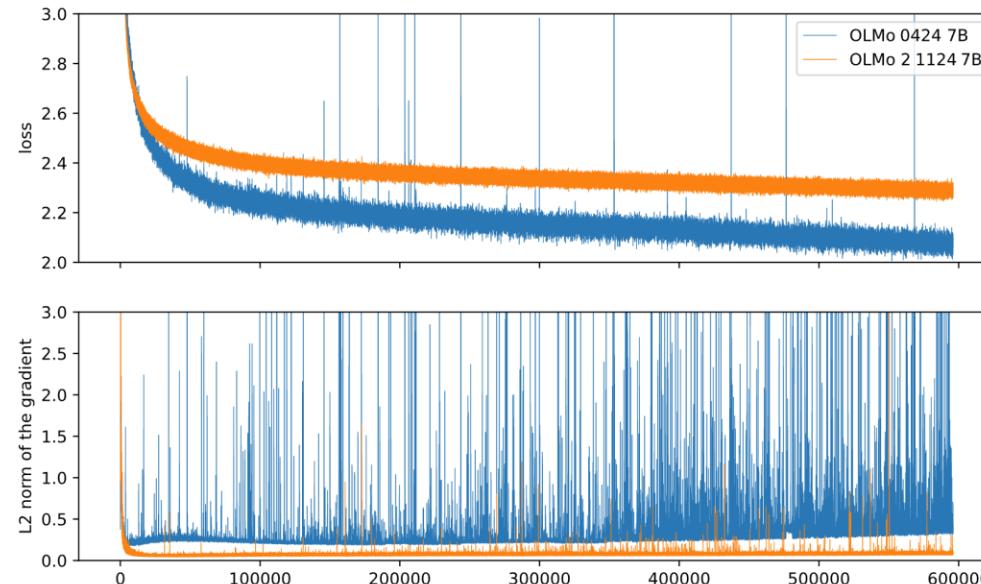
Regularization

- You still ‘regularize’ LMs but its effects are primarily on optimization dynamics

AI Name	# Year	# MLP factor	# Aspect ratio (d/layer)	# weight decay	# drop_rate
Original transformer	2017	4	85	0	0.1
GPT	2018	4	64	0.1	0.1
T5 (11B)	2019	64	43	0	0.1
GPT2	2019	4	33	0.1	0.1
T5 (XXL 11B) v1.1	2020	2.5	171	0	0
mT5	2020	2.5	171	0	0
GPT3 (175B)	2020	4	128	0.1	0.1
GPTJ	2021		146	0.1	0
LaMDA	2021	8	128		
Anthropic LM (not claudie)	2021	4	128		
Gopher (280B)	2021	4	205		
GPT-NeoX	2022	4	140	0.01	0
BLOOM (17B)	2022	4	205	0.1	0
OPT (17B)	2022	4	128	0.1	0.1
PaLM (540B)	2022	4	156		0
Chinchilla	2022	4	102		
Baichuan 2	2023	2.68	128	0.1	0
Mistral (7B)	2023	3.5	128	0.1	0
LLaMA2 (70B)	2023	3.5	102	0.1	0
LLaMA (65B)	2023	2.6875	102	0.1	0
GPT4	2023		0		
Olmo 2	2024	2.6875	128		
Gemma 2 (27B)	2024	8	100		
Nemotron-4 (340B)	2024	4	192		0
Qwen 2 (72B) - same for 2.5	2024	3.609	102		
Falcon 2 11B	2024	4	68	0.1	
Phi3 (small) - same for phi4	2024	3.5	128		
Llama 3 (70B)	2024	3.5	102		0
Reka Flash	2024		0		
Command R+	2024	2.75	192		
OLMo	2024	2.6875	128	0.1	0
Qwen (14B)	2024	2.675	128	0.1	0.1
DeepSeek (67B)	2024	2.6875	86	0.1	0
Yi (34B)	2024	2.857142	119	0.1	0
Mixtral of Experts	2024		0		
Command A	2025		0		
Gemma 3	2025	4	87		
SmolLM2 (1.7B)	2025	4	85		

Stability tricks

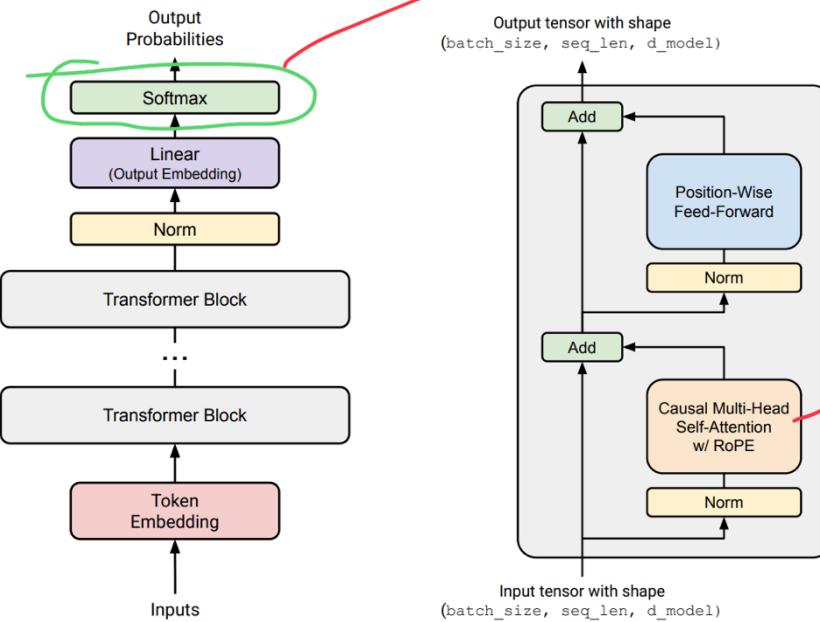
Recently, lots of attention on *stable training*



Don't train models that look like the blue curve!

Where do the issues arise? Beware of softmaxes!

Softmaxes – can be ill-behaved due to exponentials / division by zero



① Output softmax
 $\frac{1}{1+e^{-2}}$
2-loss

② Softmax in the attn. block!
↓
(or norm)

Output softmax stability – the ‘z-loss’

Recall the softmax calculation

$$\begin{aligned}\log(P(x)) &= \log\left(\frac{e^{U_r(x)}}{Z(x)}\right) \\ &= U_r(x) - \log(Z(x)) \\ Z(x) &= \sum_{r'=1}^{|V|} e^{U_{r'}(x)}\end{aligned}$$

$$\begin{aligned}L &= \sum_i [\log(P(x_i)) - \alpha(\log(Z(x_i)) - 0)^2] \\ &= \sum_i [\log(P(x_i)) - \alpha \log^2(Z(x_i))]\end{aligned}$$

[From Devlin 2014]

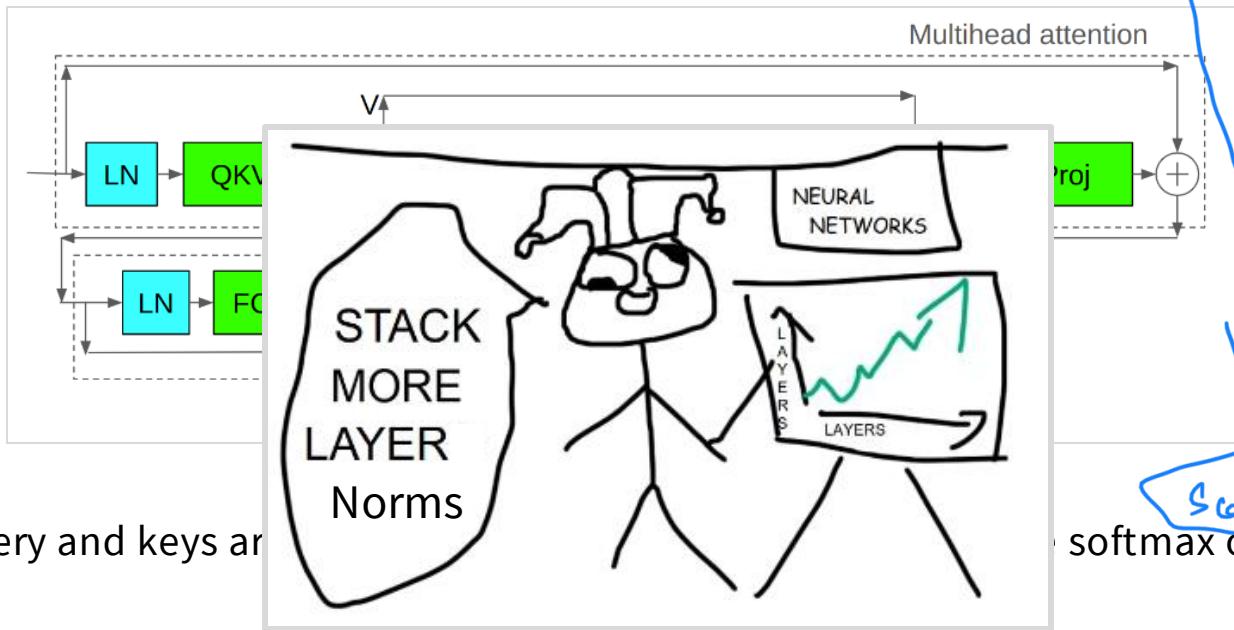
if this is close to 0, we have
much more numerical
stable softmax.

This is useful for stability! PaLM pioneered this ‘z loss’ trick.

We additionally use an auxiliary loss of $z_loss = 10^{-4} \cdot \log^2 Z$ to encourage the softmax normalizer $\log(Z)$ to be close to 0, which we found increases the stability of training.

Other examples: Baichuan 2 (2023), DCLM (2024), OLMo 2 (2025)

Attention softmax stability – the 'QK norm'



The query and keys are

Other examples: DCLM, OLMo2, Gemma 2

Originally from vision and multimodal models [Dehghani 2023, Idefcs, Chameleon]



These norms are present during inference as well.
RMS Norm trained parameters will be used to scale inputs (RMS) have layer norms for Q, K before doing BMM & Softmax!

Logit soft-capping.

Soft-capping the logits to some maximum value via Tanh

Logit soft-capping. We cap logits ([Bello et al., 2016](#)) in each attention layer and the final layer such that the value of the logits stays between -soft cap and +soft cap. More specifically, we cap the logits with the following function:

$$\text{logits} \leftarrow \text{soft_cap} * \tanh(\text{logits}/\text{soft_cap}).$$

We set the `soft_cap` parameter to 50.0 for the self-attention layers and to 30.0 for the final layer.

Prevents logits from blowing up, but also might have perf issues?

Table 4: Models perplexity with confidence interval ± 0.1 at 95% level.

bf16 baseline	<code>soft_cap</code>	<code>QKV_norm</code>	<code>QK_norm_cap</code>	<code>QK_norm</code>	<code>QK_FC_norm</code>
11.19	11.24	10.85	11.00	10.84	10.87

Attention heads

Most models don't touch the attention heads much at all with a few minor exceptions..

for inference costs! Saving.
r

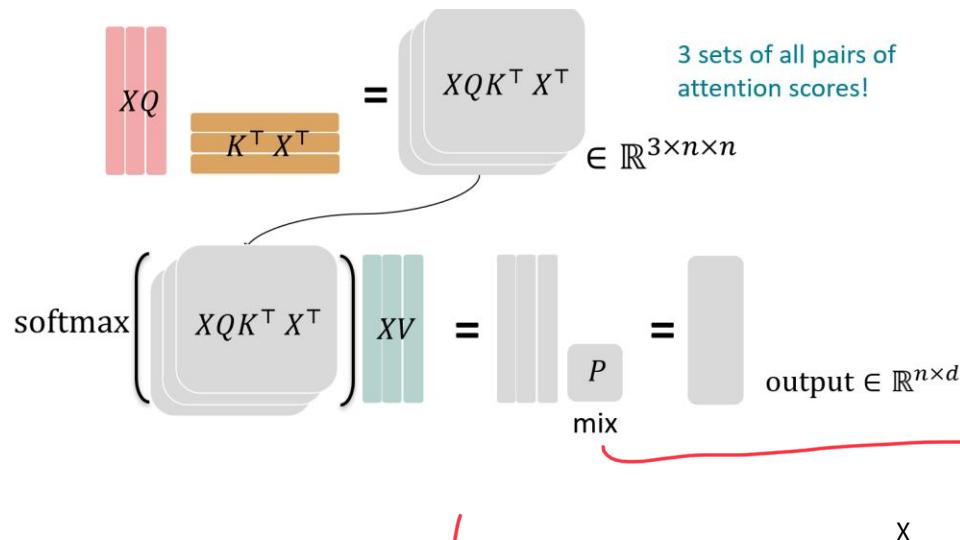
GQA / MQA: Saving inference costs by reducing the number of heads

Sparse or sliding window attention (GPT4/Mistral): restricting the attention pattern to reduce compute cost

Exotic SSM stuff (Jamba, Falcon 3, etc): not covered (sorry!)

GQA/MQA – Reducing attention head cost

Let's think about the compute involved for attention



Total arithmetic operations (bnd^2), total memory accesses ($bnd + bhn^2 + d^2$)

Arithmetic intensity is high $O\left(\left(\frac{1}{k} + \frac{1}{bn}\right)^{-1}\right)$ - we can keep our GPUs running

$b \rightarrow$ batch dim.
 $n \rightarrow$ seq. len
 $d \rightarrow$ hidden dim. length
 $h \rightarrow$ no of heads.

memory access is relatively expensive & compute is relatively cheap in GPU

X softmax projection

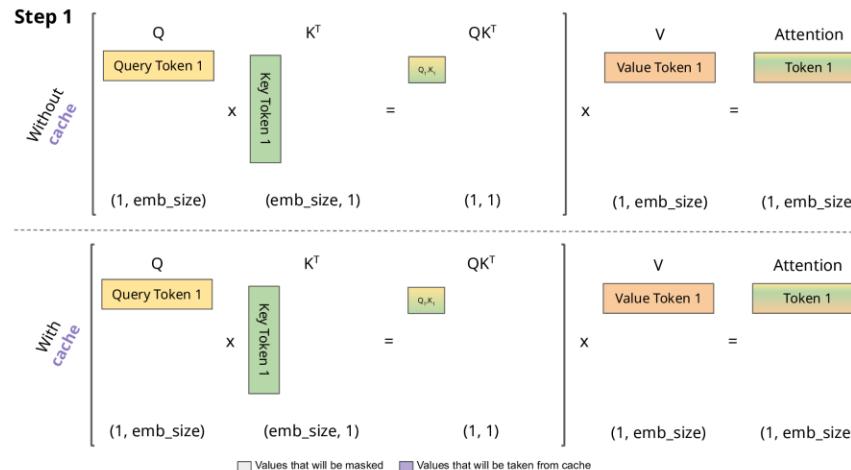
So, do lots of compute for every memory access.

GQA/MQA – Reducing attention head cost

What about the *incremental* case when we generate text?

Key difference: can't parallelize the generation process – needs to be step by step

In this case – we need to incrementally re-compute/update attention via the ‘KV cache’



GQA/MQA – Reducing attention head cost

What's the incremental arithmetic intensity?

stays same

Total arithmetic operations (bnd^2), **total memory accesses** ($bn^2d + nd^2$)

(1) *as KV-cache stores some info which needs to load/unload.*

Arithmetic intensity is not good $O\left(\left(\frac{n}{d} + \frac{1}{b}\right)^{-1}\right)$ - need large batches + short seq length
(n) or big model dimensions (d)

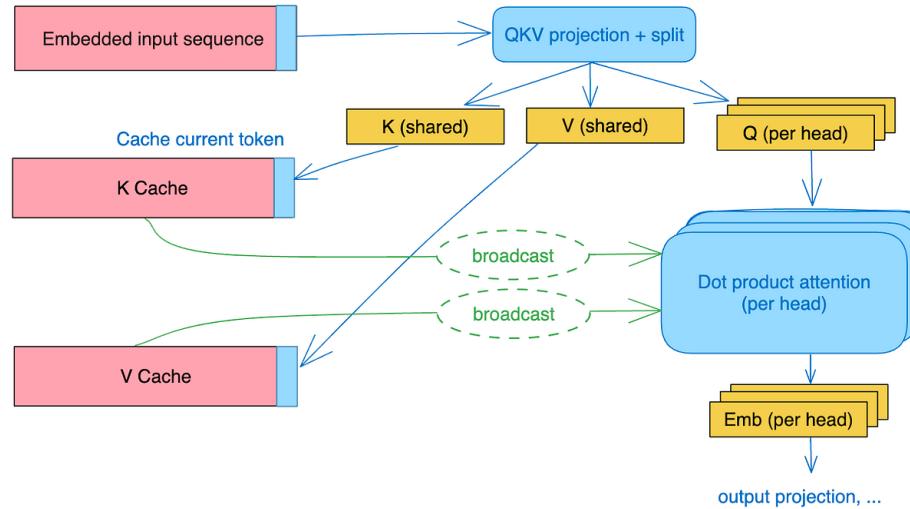
Is there some way around this? The n/d term is difficult to reduce.

MQA – just have fewer key dimensions.

Key idea – have multiple queries, but just one dimension for keys and values

have multiple heads for queries but one head for key,value.

multi-query attention



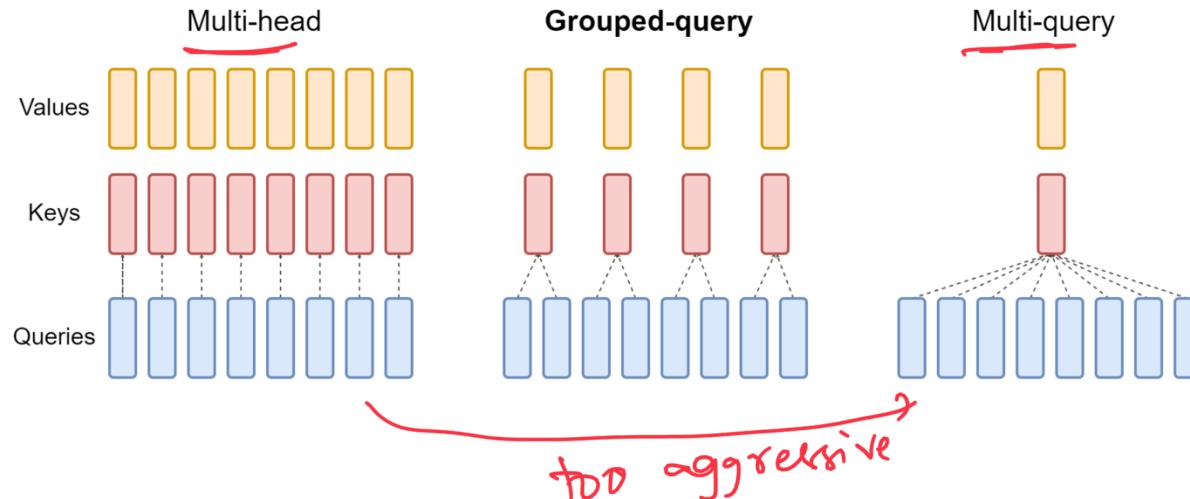
We have much fewer items to move in and out of memory (KV Cache)

Total memory access ($bnd + bn^2k + nd^2$), **Arithmetic intensity** $O\left(\left(\frac{1}{d} + \frac{n}{dh} + \frac{1}{b}\right)^{-1}\right)$

Recent extension – GQA

Don't go all the way to one dimension of KV – have fewer dims

bit slower but improves perf.



Simple knob to control expressiveness (key-query ratio) and inference efficiency

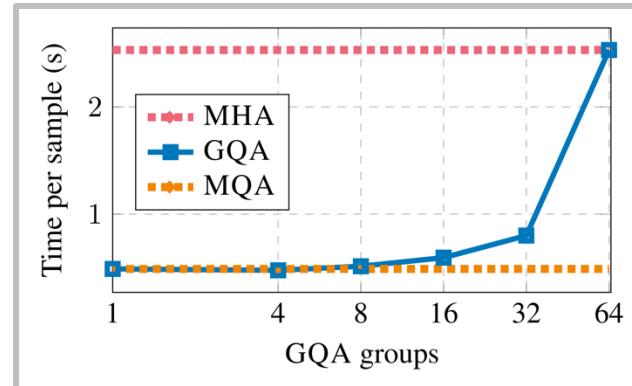
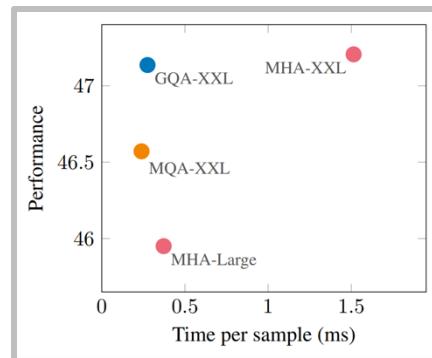
Does MQA hurt? Sometimes..

Small PPL hit w/ MQA [Shazeer 2019]

Table 3: Billion-Word LM Benchmark Results.

Attention	h	d_k, d_v	d_{ff}	dev-PPL
multi-head	8	128	8192	29.9
multi-query	8	128	9088	30.2
multi-head	1	128	9984	31.2
multi-head	2	64	9984	31.1
multi-head	4	32	9984	31.0
multi-head	8	16	9984	30.9

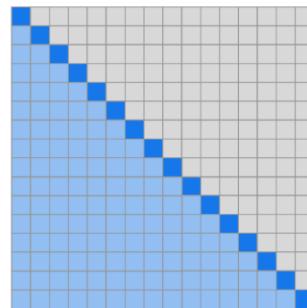
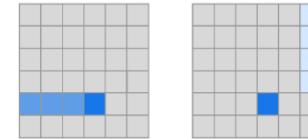
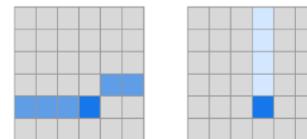
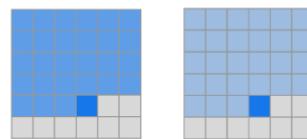
Low/no hit w/ GQA [Ainslie 2023]



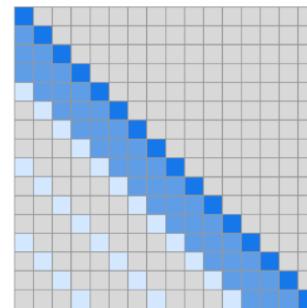
Sparse / sliding window attention

Attending to the entire context can be expensive (quadratic).

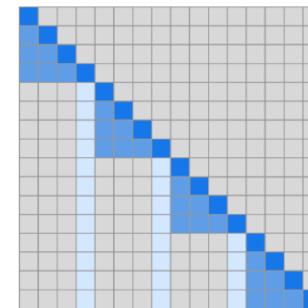
Build sparse / structured attention that trades off expressiveness vs runtime (GPT3)



(a) Transformer



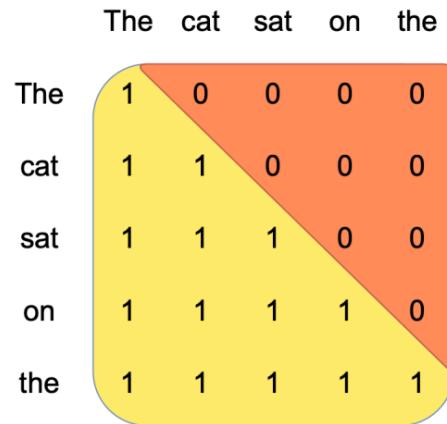
(b) Sparse Transformer (strided)



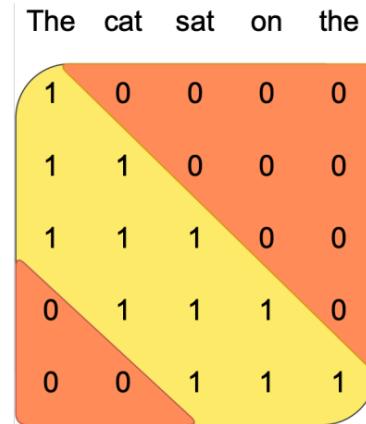
(c) Sparse Transformer (fixed)

Sliding window attention

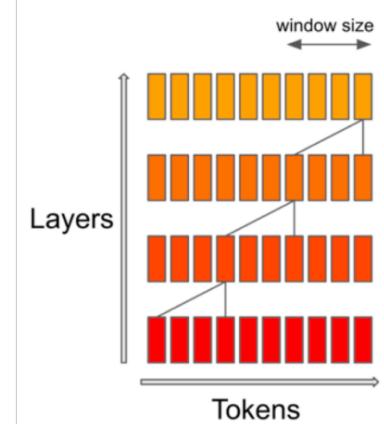
Another variation on this idea – sliding window attention



Vanilla Attention



Sliding Window Attention



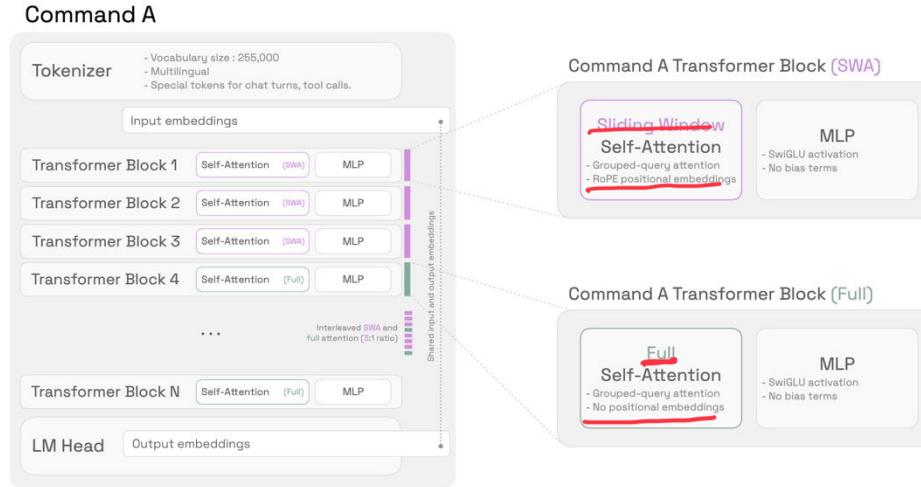
Effective Context Length

at each layer
pay attention to
only some part
rather than
entire seq
to your left!

Just use the main part of the strided pattern – let depth extend effective context (Mistral)

Current standard trick – interleave ‘full’ and ‘LR’ attention

From Cohere Command A – Every 4th layer is a full attention



Long-range info via NoPE, short-range info via RoPE + SWA.

Other models – LLaMA 4, Gemma does SWA+Full RoPE.

Recap, conclusion, etc.

Many aspects (arch, hparams) of transformers are in common across the big LMs

Aa Name	Has pa...	Link	# Year	Tokenizer type	# Vocab count	Norm	Parallel Layer	Pre-norm	Position embedding	Activations	MoE	# MLP factor	# num_layers	# model_dim
Original transformer	Yes	arxiv.org/abs....03762	2017	BPE	37000	LayerNorm	Serial	□	Sine	ReLU	□	4	6	
GPT	Yes	cdn.openai.com/res...er.pdf	2018	BPE	40257	LayerNorm	Serial	□	Absolute	GeLU	□	4	12	
GPT2	Yes	cdn.openai.com/bet...rs.pdf	2019	BPE	50257	LayerNorm	Serial	✓	Sine	GeLU	□	4	48	
T5 (11B)	Yes	arxiv.org/abs....10683	2019	SentencePiece	32128	RMSNorm	Serial	✓	Relative	ReLU	□	64	24	
GPT3 (175B)	Yes	arxiv.org/abs....14165	2020	BPE	50257	LayerNorm	Serial	✓	Sine	GeLU	□	4	96	
mT5	Yes	arxiv.org/abs....11934	2020	SentencePiece	250000	RMSNorm	Serial	✓	Relative	GeGLU	□	2.5	24	
T5 (XXL 11B) v1.1	Kind of	github.com/go...d#1511	2020	SentencePiece	32128	RMSNorm	Serial	✓	Relative	GeGLU	□	2.5	24	
Gopher (280B)	Yes	arxiv.org/abs....11446	2021	SentencePiece	32000	RMSNorm	Serial	✓	Relative	ReLU	□	4	80	
Anthropic LM (not claude)	Yes	arxiv.org/abs....00861	2021	BPE	65536			✓			□	4	64	
LaMDA	Yes	arxiv.org/abs....08239	2021	BPE	32000			✓	Relative	GeGLU	□	8	64	
GPTJ	Kind of	huggingface.co/Eh...t-j-6b	2021	BPE	50257	LayerNorm	Parallel	✓	RoPE	GeLU	□		28	
Chinchilla	Yes	arxiv.org/abs....15556	2022	SentencePiece	32000	RMSNorm	Serial	✓	Relative	ReLU	□	4	80	
PaLM (540B)	Yes	arxiv.org/abs....02311	2022	SentencePiece	256000	RMSNorm	Parallel	✓	RoPE	SwiGLU	□	4	118	
OPT (175B)	Yes	arxiv.org/abs....01068	2022	BPE	50272	LayerNorm	Serial	✓	Absolute	ReLU	□	4	96	
BLOOM (175B)	Yes	arxiv.org/abs....05100	2022	BPE	250680	LayerNorm	Serial	✓	Alibi	GeLU	□	4	70	
GPT-NeoX	Yes	arxiv.org/pdf....45.pdf	2022	BPE	50257	LayerNorm	Parallel	✓	RoPE	GeLU	□	4	44	
GPT4	OPEN	Ad	arxiv.org/abs....08774	2023	BPE	100000		□			□			
LLaMA (65B)	Yes	arxiv.org/abs....13971	2023	BPE	32000	RMSNorm	Serial	✓	RoPE	SwiGLU	□	2.6875	80	
LLaMA2 (70B)	Yes	arxiv.org/abs....09288	2023	BPE	32000	RMSNorm	Serial	✓	RoPE	SwiGLU	□	3.5	80	
Mistral (7B)	Yes	arxiv.org/abs....06825	2023	BPE	32000	RMSNorm	Serial	✓	RoPE	SwiGLU	□	3.5	32	

Major differences? Position embeddings, activations, tokenization