Model habs after defining a useaseq most of the time Cheering from a series  I proteined models, are belter than training one  I feel list of Available pre-trained models, rapid be fromed an Hugging to  hub, If Huh,  Item Are UMs Pre-train?  A high level look:  1.37 of eigin!  Teat! I Perform 17 the 1974  Teat I I Perform 1852  Teat I I Perform 1852  I render 1974  The 1236  Uscabrilary  See .18. PB  I unstructured  data  Training toward on 11st types: each type of 11m structure, is best for a cortain seed to the served on the trained of the free form of the served of training models.  The tracker togething (MLM)  Objective Reconstruct text ("denoising")  The tracker togething (uses both the Gaster) to before and the to provide masker  Teaching (uses both the Gaster) to before and the to provide masker  Teaching (uses both the Gaster) to before and the to provide masker	11 1/1 11	19.	& The time Changing	from a	Spries
How see 11 st flut 1. 3i of origin!  Test 1. 1	Model habs all	er detining a usecase, most a	, me may one	provi a	Jane,
How see 11 st flut 1. 3i of origin!  Test 1. 1	of preticined in	rodels, are better than train	ny one Con 1	, ,	
How see 11 st flut 1. 3i of origin!  Test 1. 1	1 full list	of Available pre-trained mo	Jels, per be trune	/ ca to	ugging
A high level look  1. 37. S. criginal  Test 1. 1 27 1-00 57,  Test 1. 1 28 1-11 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	hub, Ff +	Juh,			
A high level look  1.32 of ciginal  Token string Token ID Embelling  Text [ ]					
A high level look  1.32 of ciginal  Token string Token ID Embelling  Text [ ]			•		*
A high level look  1.32 of ciginal  Token string Token ID Embelling  Text [ ]	14-1 Aco 11 Mg	Prestain 9			
Test	Jew Me Cons	•			
Test   -	11.11.11				
Test [-1]   Test   Test	1 high revel look	127 8 1	Token string	TokenID	Embedd
Test []   Ital   I   I   I   I   I   I   I   I   I			1-The	37	1-005
Test 1-1   1236    Test 1-1   1236    The Test   12	1ex 1 3	Tolers .	'Toucher'	11852	
South the structure of the structure is best for a cortain so of teasts, and is the trained of the structure	1ent 1-~   5	->   Test   ->   L M	+	19741	
South Tell  South Tell  South Tell  South To the structure of the structure is best for a cortain so of tests, and is the trained differently:  A) Auto encoding models (Encoder only LLM)  Masked longuege modeling (MLM)  Masked Example models:  BERT  ROBERTA  Objective: Reconstruct text ("denoising")  Min   tendrer Lytisks ) the (stouents)	Test 1 1			1736	-
CB - TB - PB  CI unstructured  date.  Itaining town on 11 of Types each type of 11 m structure, is best for a cortain's  P tasks, and is lie trained differently:  A) Auto encoding models (Encoder only 11 m)  Masks   Ionguage modeling (MLM) - named entity recognitive teacher   teacher   teacher   the [stodenty    Masks   Example models:  Masks   Example models:  PART    Objective: Reconstruct text ("denoising")  The   teacher   texts   the [stodents    Differently    Objective: Reconstruct text ("denoising")	1	Vial-Tob	1me-		
data  Jeaning town on 11 of Types each type of 11m structure, is best for a cortain s  - I tasks, and is the trained differently:  A) Auto encoding models (Encoder only 11m)  Masked longuage modeling (MLM)  - named entity recogni  - word classification  Into teacher teaching the Istodenty  Mask  Example models:  - BERT  - ROBERTA  Objective: Reconstruct text ("denoising")  The teacher exists the Istodents  Didirectional Context	100	Pro La	Voc	bolony	
Juning toxed on 11st Types: each type of LLm structure, is best for a cortain's  Plasks, and is the trained differently:  A) Auto encoding models (Encoder only LLM)  Musked longuage modeling (MLM)  -named entity recogni  -wow does in cation  Whe teacher teaching the stodenty  Example models:  DERT  ROBERTA  Objective: Reconstruct text ("denoising")  Will teacher wisks the fistwests  Differentiated Context	GB-TB-PB		A Company of the Comp		
A) Auto encoding models (Encoder only LLM)  Musical language modeling (MLM)  Maskal language m	f unch 1				
A) Auto encoding models (Encoder only LLM)  Musical language modeling (MLM)  Musical language modeling (MLM)  -named entity recognitive teacher teacher teacher the stodenty  Masky  Encoder and Encodes:  BERT  ROBERTA  Objective: Reconstruct text ("denoising")  The tencher encodes  Didirectional context	Cronstitutived				
A) Auto encoding models (Encoder only LLM)  Musical language modeling (MLM)  -named entity recognitive teacher teacher teaching the stodenty  Encope models:  Masky  Encope models:  BERT  ROBERTA  Objective: Reconstruct text ("denoising")  The tencher teacher text ("denoising")	_data				
A) Auto encoding models (Encoder only LLM)  Musical language modeling (MLM)  Musical language modeling (MLM)  -named entity recogni  -word classification  Whe teacher teaching the stodenty  Mask  Encople models:  BERT  ROBERTA  Objective: Reconstruct text ("denoising")  The tenched wasks the stodents  Didirectional Context					
A) Auto encoding models (Encoder only LLM)  Musical language modeling (MLM)  Musical language modeling (MLM)  -named entity recogni  -word classification  Whe teacher teaching the stodenty  Mask  Encople models:  BERT  ROBERTA  Objective: Reconstruct text ("denoising")  The tenched wasks the stodents  Didirectional Context				0	
A) Auto encoding models (Incoder only LLM)  - sentimental anxlys  - manked longuage modeling (MLM)  - named entity recogni  - word classification    The   teacher   teaching   the   stodenty    Enough models:    BERT   ROBERTA    Objective: Reconstruct text ("denoising")    The   teacher   wask   The   stowents   bidirectional Context	Training based on	Un Types each type of I	'Im structure, is best	fora	certain
A) Auto encoding models (Incoder only LLM)  - sentimental anxlys  - manked longuage modeling (MLM)  - named entity recogni  - word classification    The   teacher   teaching   the   stodenty    Eacomple models:    BERT   ROBERTA    Objective: Reconstruct text ("denoising")    The   teacher   wask ) The   stowents   bidirectional Context	Training based on  Plasks, and	11 of Types each type of I is Pretrained differently	'Im structure, is best	fora	certain
Masked longuage modeling (MLM)  -named entity recogni  -word degsi hication    The   teacher   teaching   the   stodenty     Mask   Eaconple models:    BERT     ROBERTA     Charles   Construct text ("denoising")     Min   tenched   Mask   The   stowerts     Divined ignal Contest	Franks, and	un Types each type of I is Pretrained differently	:		
The teacher teaching the stodenty  Masky  Encomple models:  BERT  Construct text ("denoising")  The tenchen (Mask) the students  Didirectional Context	-Frasks, and	is Vietraised differently	Good vse	Coers:	
Ithe teacher teaching the stodenty  Masky  Eaconple models:  BERT  ROBERTA  Objective: Reconstruct text ("denoising")  The Fencher wasky the students  bidirectional Context	-Frasks, and	is Vietraised differently	(Good vse	Ceses:	anely
Description of the tender wask the students bidirectional Contest	-Frasks, and	y models (Encoder only 14)	Good vse	Ceses:	anely
Description of the tender wask the students bidirectional Contest	-Frasks, and	y models (Encoder only 14)	('ULU) -name	Ceses: imental	anely
Objective: Reconstruct text ("denoising")  The tendren exists the Estwerts  bidirectional Context	-Frasks, and	y models (Encoder only 11)  Musical language modeling	('ULU) -name -word	Ceses: imental	anely
Objective: Reconstruct text ("denoising")  The tendred wask) The students  bidirectional Context	-Frasks, and	models (Encoder only 11)  Musical longuage modeling  The teacher teaching the	('MLM) -name  [stodenty]	Ceses: imentations desifica	anely lecogn
Objective: Reconstruct text ("denoising")  The Fencher wask) The Istouents  bidirectional Context	-Frasks, and	models (Encoder only 11)  Musical longuage modeling  The teacher teaching the	(MLM) -name -word  [stodenty]  Example	Ceses: imentations desifica	anely lecogn
The tendred wask the students bidirectional Content	-Frasks, and	models (Encoder only 11)  Maskad longuage modeling  The teacher teaching the  Masky	(MLM) -name -word  [stodenty]  Eacomple  BER	Ceses: imental Tontity chasifica	anely lecogn
The tendred wask the students bidirectional Context	-Frasks, and	models (Encoder only 11)  Maskad longuage modeling  The teacher teaching the  Masky	(MLM) -name -word  [stodenty]  Eacomple  BER	Ceses: imental Tontity chasifica	anely lecogn
The tenched wask the students bidirectional Context	-Frasks, and	models (Encoder only 11)  Maskad longuage modeling  The teacher teaching the  Masky	(MLM) -name -word  [stodenty]  Eacomple  BER	Ceses: imental Tontity chasifica	anely lecogn
bidirectional Context	A) Auto encoding	y models (Encoder only 11)  Masked longuage modeling  The teacher teaching the  Masky  Encoder only  Lu  Lu  Lu  Lu  Lu  Lu  Lu  Lu  Lu  L	('ULM) -sent -sent -word  [stodenty]  Eacomple -ROBI	Ceses: imental Tontity chasifica	anely lecogn
bidirectional Context	A) Auto encoding	y models (Encoder only 11)  Masked longuage modeling  The teacher teaching the  Masky  Encoder only  Lu  Lu  Lu  Lu  Lu  Lu  Lu  Lu  Lu  L	('ULM) -sent -sent -word  [stodenty]  Eacomple -ROBI	Ceses: imental Tontity chasifica	anely lecogn
teaching (use's both the Content (to kens) before and after to predict wasked	A) Auto en coding	y models (Encoder only 11)  Maskad longuage modeling  The teacher teaching the  Masky  enooder-only  LLM  jective: Reconstruct text ("o	(MLM) -name -word  [stodenty]  Enough  ROBI  enoising 1)	Ceses: imental Tontity chasifica	anely lecogn
before and after to predict wasked	A) Auto encoding	y models (Encoder only 11)  Maskad longuage modeling  The teacher teaching the  Masky  enooder-only  LLM  jective: Reconstruct text ("o	(MLM) -name -word (Stodenty)  [stodenty]  Enough -ROBI  Towlerts  Towlerts	Ceses: imental Jentity chasifica  models T FRIA	anely lecogn
	A) Auto en coding	models (Encoder only It.  Musical language modeling  The teacher teaching the  Masky  enoder-only  LLM  itho teacher wasky the for	(MLM) -name -word  [stodenty]  Enoising 1)  towerts  bidirectional Company  Towerts	Ceses: imental Identity clossifica models T RTA	anely lecogn tion

B) Autoregressive models: (Decoder-Only)	
Causal language modeling (CLM)	
Causal language modeling (CLM) Akh. full a a	Good use cases:
	Pert generation
7 The teacher 9	
	Ocepant on the model
( Decoder Only)	
model	Example models:
	- GPT
Objective: Peedict next token	-BLOOM
The reached teachers	
idirectional Contest	
earing the modelarly uses	
earing the modelarly uses evious tokens to generate w taken)	
w token 1	
C) sequence-to-sequence models: (encodes spen Corruption	
The Yeard (mosky Knowsky Bludent)	Good use-cases
The Hoscher 1' ( on 1 students)	-Translation
sentient token	-Summerization
	- avestion answering
Enc. Dec j	
	Example madels:
Objective: reconstruct span	76
1 - Construct span	-BART
(2n) teaches the	
Predictive	
tokens	
clips	тм —

In Tale On II IN was upon	4 butes for each	our meter	
to run an LLU, we need	A STATE OF THE PARTY OF THE PAR		mizer (e& byte)
to train it 20 bytes per po	rameter goriginal 4	bytes Gradients	(+4 byte)
			of (+ Shyles)
So approx for a 1B pava moo	lel: 46B@32611.		7
	to store mode	5 8000	@ 32 bit
			ercision
What we am do to store	and train models	on to the	in the mode
a smaller GPU Ran ?			
•			125.7
Quantization: the process	to projecting pura	meters to smaller	drypes/sizes to
reduce The short train	Size of LLM. Th	= might dtypelo	store paras is 1130
which requires 4 bytes,	but we try to pr	ged it's valve t	smaller dtypes such
as FP16, BFloat 16 or	IN18:		Memory needed
Bits	Exponent	Fractica	to store on Valve
FP 32 32	-8	23	4 byte
FP16 16	5	10	2 byte
BFLOAT16 16	8	7	2 byte
INT8 8	1 -/-	7	1 byte
INT8 8 best structure for Quant	zation for now cles	s Joss Man FP1	6)
	1,	0	
- So we Quantize, to redoc Cost of Percision	e memory needed }	or storing and train	ing LM, at the
Cost of Percision			
			0
- Quantization_aware Praining	(all) learnes the au	entization 5 Caling	tactors during training
P. ( Flan Tr.)	vained by QAT on	BYLUALIG	
for instance, google T5 is			
for instance, google T5 is			
for instance, gogle 75 is			
for instance, gogle T5 is			
for instance, google 75 is			
- Quantization_aware training for instance, google T5 is			

model size: l'arameter sie		All All
Scaling laws and Compute op	timal models:	
The final goal of training any		Model loss. Optima!
To do so, we can incresse: 50 20 20 Mar	atuset size Chumber of	tohens) models)
2, Ma	del size crumber of pa	aneters)
but we also have compute bud	get (BPUs, training time	Cost) as Constraints.
-we use petaflops 15-dong as a	unit to show the Computer by a floating point aps po	budget need for training each max exformed at rate of 1 peta FLOP1S 1100; for one day
Coequal to 2 N	Avidia V100s or 2 Nvidia	71005
How is the relation of each	Variable in model loss?	
and model size liked,  loss decreases with improved C!	Compute to budget	Hand yet again, increuse in model size ends op with lower loss
-CB		lower loss
	model 3	Con the second s
Pata	set model	M3
1 1 2 1	Size	
At again, by fixing two		
decreases with larger DS		
occueases with larger v		
And is there any relation/opti	mal size and relation	ship between them?
In Chinchilla paper, it is show.	that there is a = 20	1x relation between model sice
and Data set size; to have an	optimal model with si	re ct quB parameters, we need
2000 tokens: Madel # of para	Compute aptimal	NO COCCI
Chinchilla 70B	~ 1.4 T	1.4 B
LLaMa-65B 65B	~1.37	1.4 3
GPT-3 175B	~ 3.57	300 B 7
OPT-175B 175B	~ 3.5T	1800 under trained
BLOOM 176B	~2375 I	3500

So in Contrast of what we thought, just having bigger models wan't end us up with better performance A in one new paper, it is shown that not only chichilla rule is correct, but having a higher quality dataset also may endup in better performance models, with smaller destagets (~ 5x) than chinchilla models

it is achived by (running a classifier on data set and hand picking related texts for models

- separating more useful texts by kegword search. If the domain we want to work on, her an specific vocabulary and lenguage structure which is uncommon, Prestrain is necessary. Specific terms, vocab and language structure. Another great exemple is Bloomberg GPT, which stricktly tries to apply chichilla scaling rule, trained on 511 Public and Private finantial data Cit has specific rules for Compute budget and model structure data