The “Deep Learning for NLP” Lecture Roadmap

Lecture 10.5: Finetuning

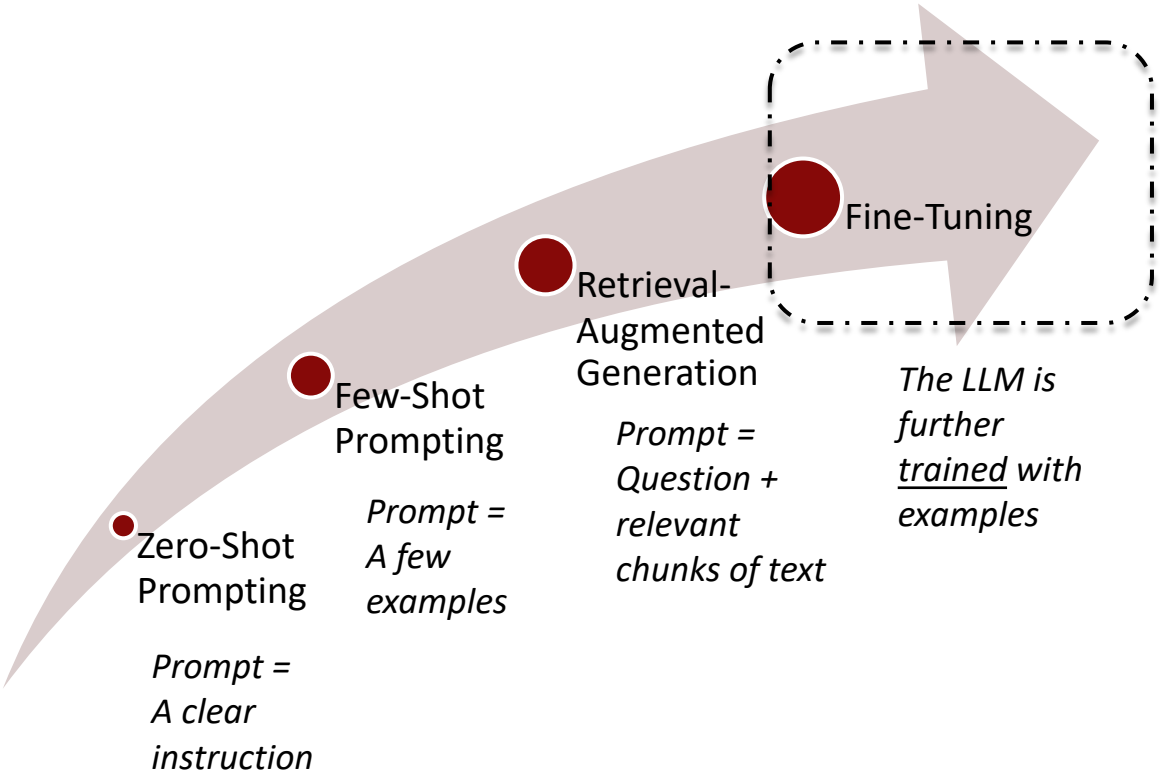
~~Lecture 5: Text Vectorization and the Bag-of-Words Model~~
~~Lecture 6: Embeddings~~
~~Lecture 7: Transformers – (1/2)~~
~~Lecture 8: Transformers – (2/2)~~
~~Lecture 9: LLMs (1/2)~~
~~Lecture 10: LLMs (2/2)~~

MIT MANAGEMENT SLOAN SCHOOL

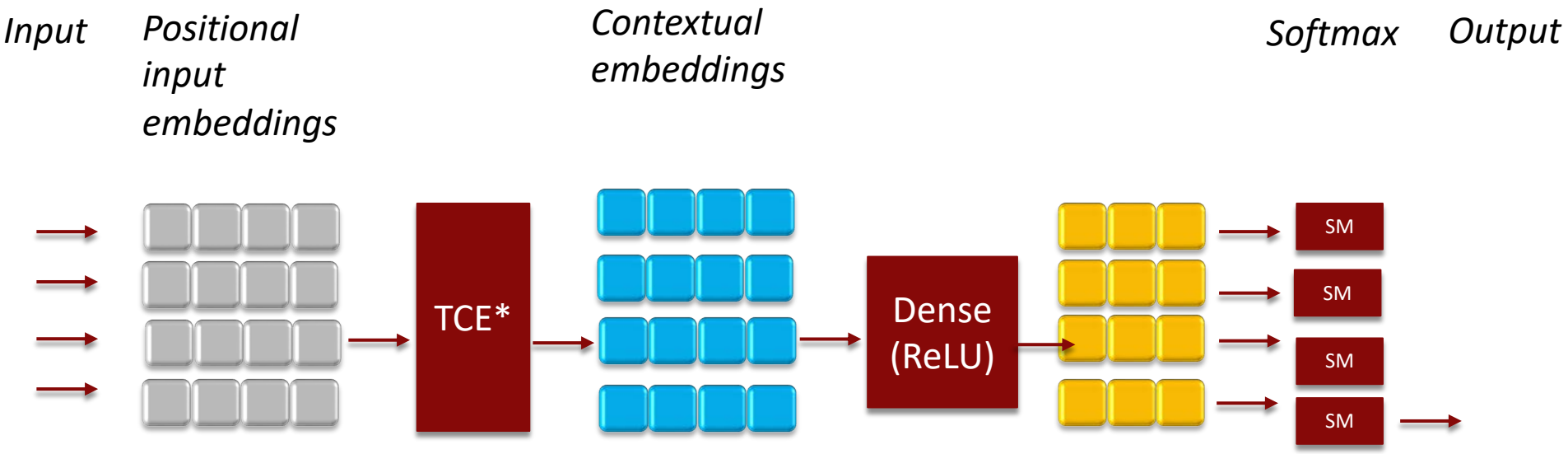
15.S04: Hands-on Deep Learning
Spring 2024

Farias, Ramakrishnan

Let’s look at Fine Tuning next



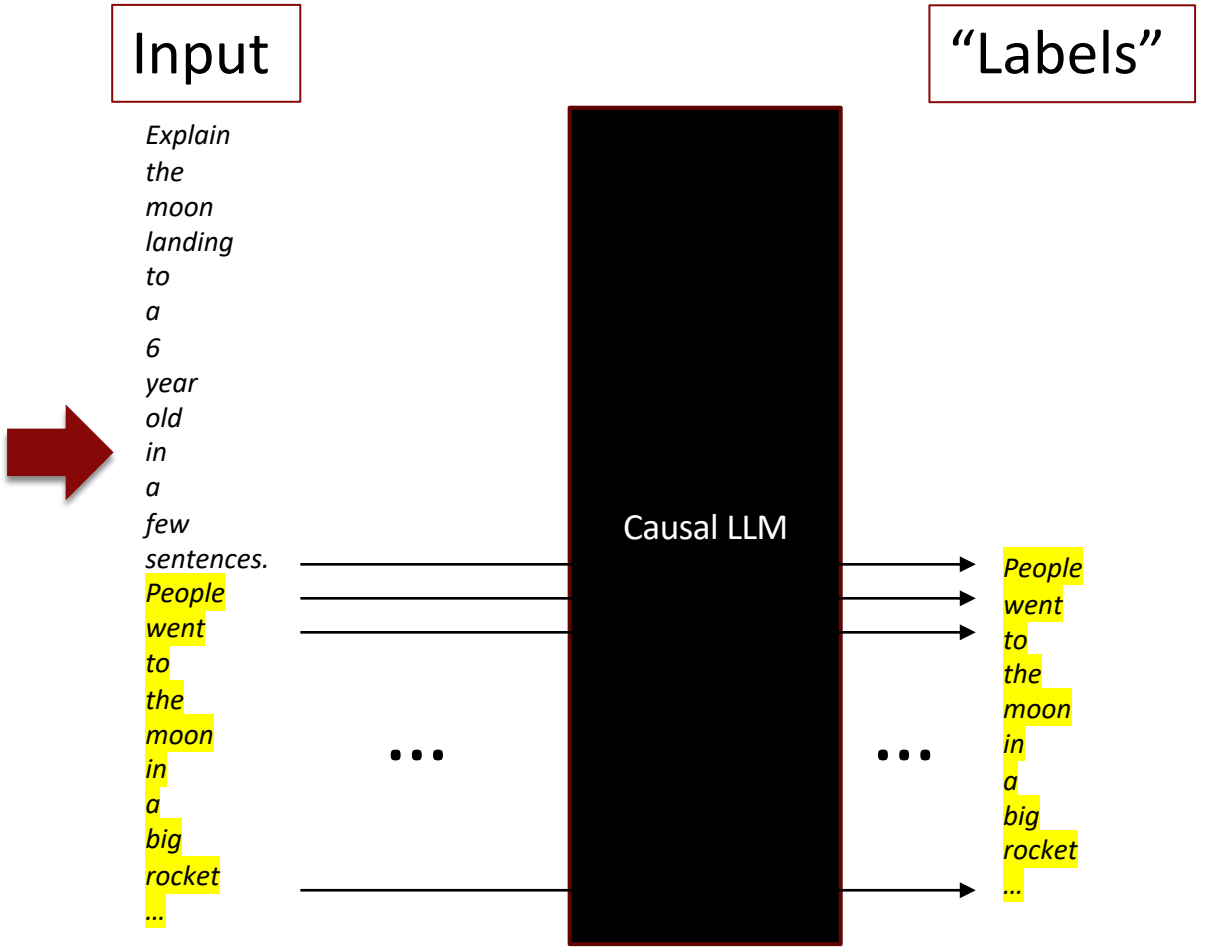
In Fine-Tuning, we take a causal LLM (like GPT) and …



*Transformer Causal Encoder stack

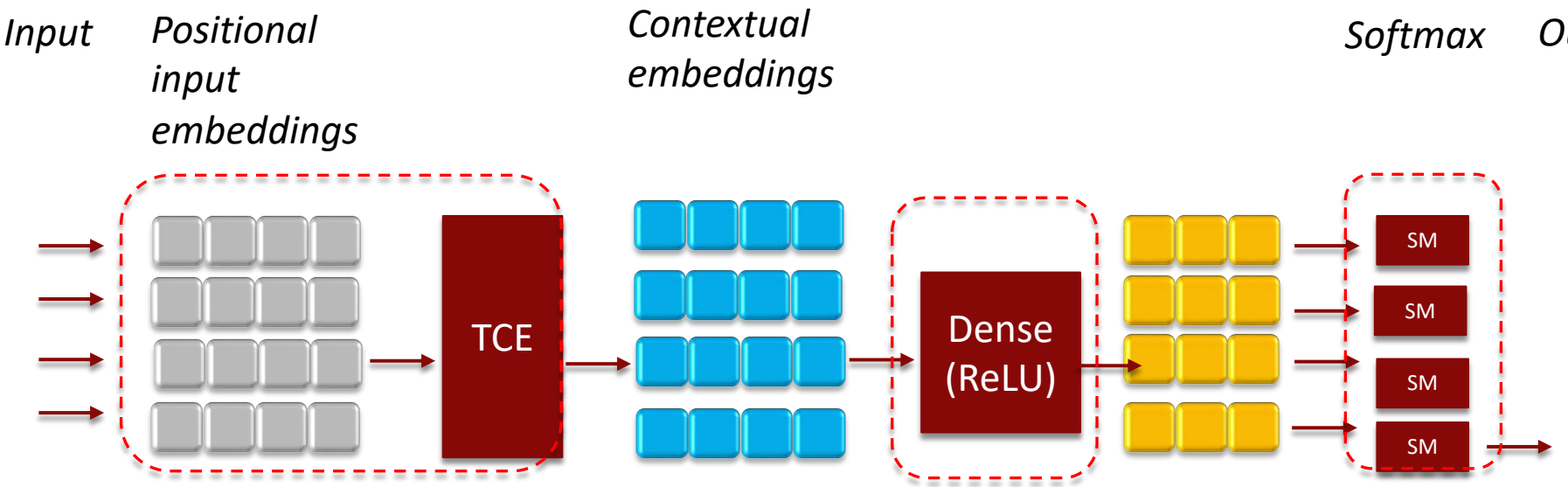
… train it further on domain-specific input-output examples …

Input	*Explain the moon landing to a 6 year old in a few sentences.*
Desired output	People went to the moon in a big rocket, walked around and came back to Earth. They took pictures of what they saw and sent them back so we could all see them.



* This is essentially the Supervised Fine-tuning step we discussed earlier to transform GPT-3 to GPT-3.5

… and update the weights



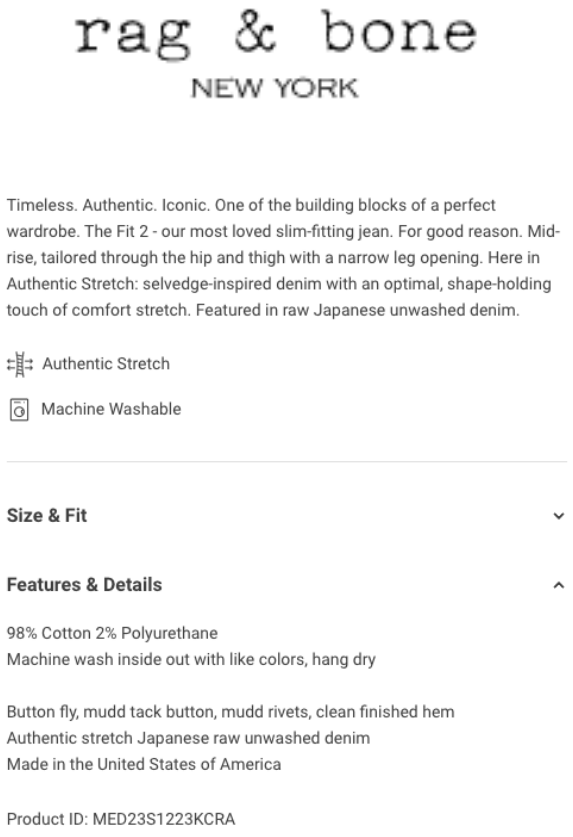
1. *Positional embeddings*
2. *Stand-alone embeddings (unless pretrained and Trainable=False)*
3. *Matrices A^K, A^Q, A^V for each attention head (inside TCE)*
4. *Layer norm scale and bias parameters (inside TCE)*
5. *Weights in Feed-Forward layers (inside TCE)*
6. *Weights in Dense layers outside TCE*
7. *Weights in final Softmax layer*

For small causal LLMs (like GPT-2), this is possible but for larger LLMs, we will face computational challenges.

We will explain the nature of these challenges and describe an alternative approach that addresses these challenges.

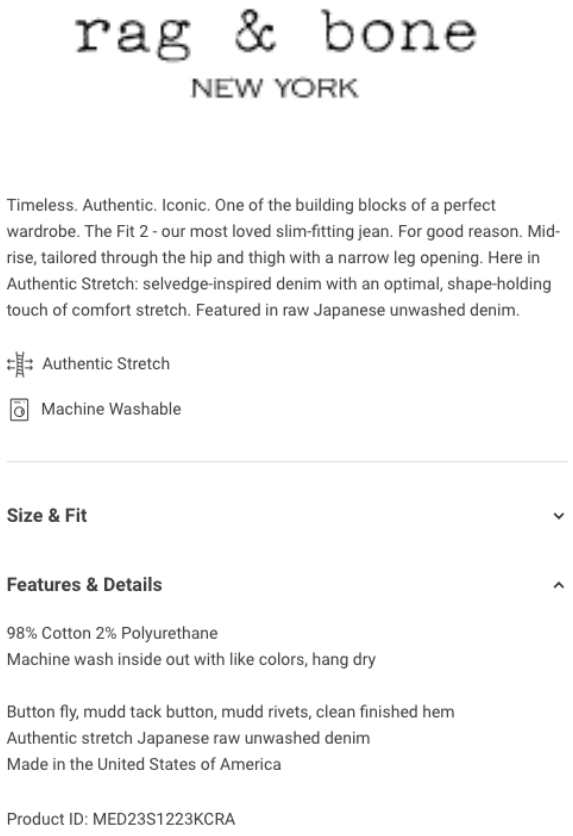
Bit first… why would we want to do this?

‘Write a positive review of the following product’ What do you think?



Timeless. Authentic. Iconic. One of the building blocks of a perfect wardrobe. The Fit 2 - our most loved slim-fitting jean. For good reason. Mid-rise, tailored through the hip and thigh with a narrow leg opening. Here in Authentic Stretch: selvedge-inspired denim with an optimal, shape-holding touch of comfort stretch. Featured in raw Japanese unwashed denim. Authentic Stretch. Machine Washable. Mid rise slim fit, tailored through hip and thigh

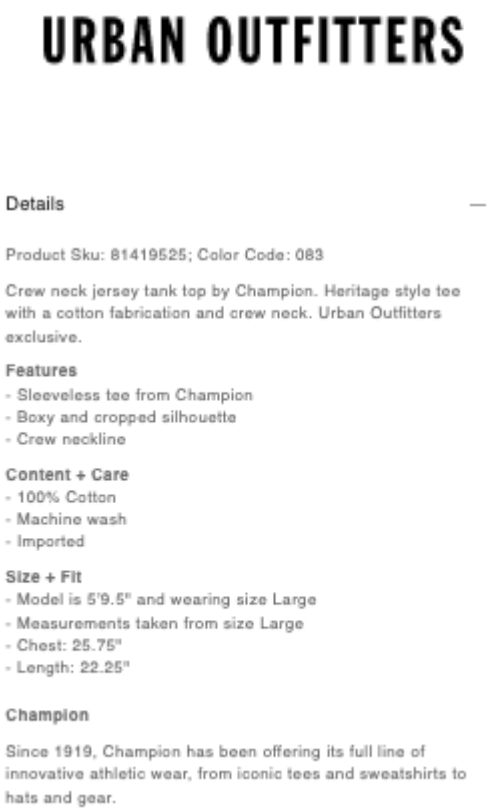
‘Write a positive review of the following product’ What do you think?



Timeless. Authentic. Iconic. One of the building blocks of a perfect wardrobe. The Fit 2 - our most loved slim-fitting jean. For good reason. Mid-rise, tailored through the hip and thigh with a narrow leg opening. Here in Authentic Stretch: selvedge-inspired denim with an optimal, shape-holding touch of comfort stretch. Featured in raw Japanese unwashed denim. Authentic Stretch. Machine Washable. Mid rise slim fit, tailored through hip and thigh

This is Alpaca, running off of Llama 7B. Reads like Marketing copy.

‘Write a positive review of the following product’ What do you think?



URBAN OUTFITTERS

Details

Product Sku: 81419525; Color Code: 083

Crew neck jersey tank top by Champion. Heritage style tee with a cotton fabrication and crew neck. Urban Outfitters exclusive.

Features
- Sleeveless tee from Champion
- Boxy and cropped silhouette
- Crew neckline

Content + Care
- 100% Cotton
- Machine wash
- Imported

Size + Fit
- Model is 5'9.5" and wearing size Large
- Measurements taken from size Large
- Chest: 25.75"
- Length: 22.25"

Champion

Since 1919, Champion has been offering its full line of innovative athletic wear, from iconic tees and sweatshirts to hats and gear.

Champion's Crew Neck Jersey Tank Top is a great addition to any wardrobe. The boxy and cropped silhouette is flattering on all body types, and the cotton fabrication is soft and comfortable. The crew neckline is classic and timeless, and the sleeveless design makes it perfect for warmer weather. The Heritage style tee is an exclusive from Urban Outfitters, and the 100% cotton fabrication and machine washability make it easy to care for. The boxy and cropped silhouette is flattering on

Same story – reads like marketing copy.

Lets get better at Fake Reviews


```
{
    "instruction": "Write a positive review for the following product",
    "input": "The Workwear Pocket Long Sleeve T-shirt.Wear it as a layer, wear it by itself, wear it with pride. It's as rugged and dependable as your favorite hammer
    offering 6.75-ounce, 100% cotton jersey knit, a generous left-chest pocket, side-seamed construction to minimize twisting and crew neck comfort. Made of 6.75-ounce,
    100% cotton jersey our long sleeve pocket t-shirt makes for perfect workwear layered or worn alone. It features a crew neck with a generous left chest pocket and
    side-seamed construction to minimize twisting. Ash fiber content is 99%/1% polyester. Heather gray fiber content is 90% cotton/10% polyester. 100% Cotton Imported
    Machine Wash Long-sleeve t-shirt in midweight jersey featuring spade patch pocket at chest with logo tag Ribbed cuffs Original fit Package
    Dimensions:\n                    \n12.4 x 7.6 x 1.8 inches Shipping Weight:\n                    \n1 pounds (View shipping rates and policies)",
    "output": "The Best! Carharrt t-shirts are my husband's favorite.  They fit well, the fabric is very substantial, they wear well, and he especially likes the pocket.
    He wears these alone, with a light t-shirt under, or under a long sleeve button down Carharrt work shirt."
},
```

```
{
    "instruction": "Write a neutral review for the following product",
    "input": "These overalls are made of 12-ounce, firm-hand 100 percent ring-spun cotton duck and have adjustable front-elastic suspenders, a multi-compartment bib
    pocket, a zipper fly, two large lower-front pockets a left-leg hammer loop, double-front knee construction with cleanout bottoms that accommodate knee pads, two
    reinforced back pockets, triple-stitched main seams and metal rivets at vital stress points. Care-Do not bleach Our duck bib overall is made of heavy 12-ounce, cotton
    duck. It features a variety of tool and utility pockets as well as a hammer loop. Double knees can accommodate knee pads and feature cleanout bottoms. Back pockets
    are reinforced for durability. 100% Cotton Imported Machine wash warm with like colors, Do not bleach, Tumble dry medium remove promptly Bib overall in 12-ounce
    heavyweight cotton duck featuring hammer loop and reinforced paneling at knees Side utility pockets and bib pocket with logo patch double knees Kindly refer last
    Image for sizing details Multiple tool pockets Package Dimensions:\n                    \n15.6 x 6.3 x 3.5 inches Shipping Weight:\n                    \n2.95 pounds
    (View shipping rates and policies)",
    "output": "... a little leary about a new pair being as good because. I have a 15 year old pair of these that got a hole burned through them when using a cut off saw
    and was a little leary about a new pair being as good because... well, I am in my 50's and have seen the vast majority of consumer/professional goods turn to crap. So
    far so good."
},
```

```
{
    "instruction": "Write a negative review for the following product",
    "input": "The iconic, timeless chuck taylor all star sneaker. Perfect in its simplicity since 1917. Lightweight, breathable canvas construction. Vulcanized rubber
    sole delivers durable traction. An ortholite insole cushions each and every step. Versatile, fashionable, and they keep looking better every time you wear them.
    Ortholite is a trademark of O2 partners, llc. 100% Textile Imported Rubber sole Shaft measures approximately low-top from arch Lace-up, low-top sneaker OrthoLite
    insole for cushioning Medial eyelets for airflow Canvas upper Product Dimensions:\n                    \n5 x 5 x 0.7 inches Shipping Weight:\n                    \n12
    ounces (View shipping rates and policies)",
    "output": "I love the classic look of Chuck Taylors I love the classic look of Chuck Taylors, but wow, these are legitimately the most uncomfortable shoes I've ever
    worn. I wore them recently to my work where I'm on my feet for 10 hours and am moving around a lot and lifting stuff, climbing up and down ladders, and by the end of
    the night I was limping and my feet were throbbing. I have another pair of Chuck Taylors, mid tops, and those have some padding in the sole and I can wear them all
    night and feel fine. I guess I'm going to have to get some inserts for these if I want to wear them longer than an hour."
},
```

The Llama 2 family of open LLMs are widely used for fine-tuning

Large language model

Llama 2: open source, free for research and commercial use

We're unlocking the power of these large language models. Our latest version of Llama – Llama 2 – is now accessible to individuals, creators, researchers, and businesses so they can experiment, innovate, and scale their ideas responsibly.

Download the model

https://llama.meta.com/llama2

Let’s first understand how “hard” it is to build the biggest model in the family: Llama-2-70b

Large language model

Llama 2: open source, free for research and commercial use

We're unlocking the power of these large language models. Our latest version of Llama – Llama 2 – is now accessible to individuals, creators, researchers, and businesses so they can experiment, innovate, and scale their ideas responsibly.

Download the model

https://llama.meta.com/llama2

How hard is it to train Llama-2-70b?

- The model is gigantic
  - 70 billion parameters x 2 bytes per parameter x ~3-6 = 420-840GB
    - 2 bytes/parameter assuming we use “fp16” (i.e., 16-bit numbers)
    - 3-6x multiplier for each parameter since we need to store gradient, optimizer states etc. in addition to the weights themselves, and some require higher precision (more on this later)

- An A100 (or H100) GPU has 80GB of RAM and so we need between 6 and 11 GPUs to accommodate 420-840GB

How hard is it to train Llama-2-70b?

- Llama-2-70b was trained on 2 trillion (2×10^{12}) tokens
- An A100 GPU can (optimistically) process just about 400 tokens/GPU per second
- So 11 GPUs will take $2 \times 10^{12}/(11 \times 400)$ seconds which is about 5,261 days. A bit long ☺
- Let's say we want to do it in about a month.
  - With 2048 GPUs, we would need about 28 days
  - A simple cost estimate (at $2.5/ GPU-hr) for this training run is $4M
  - We would expect the actual cost to be a lot higher since it takes multiple runs to get things right

To fine-tune Llama-2-70b with fewer resources, we need to do two things

- Reduce the size of the dataset
- Reduce the memory required so that we can process the data using fewer GPUs (ideally, just one GPU)

To fine-tune Llama-2-70b with fewer resources, we need to do two things

- ✅ Reduce the size of the dataset: We are in luck here! Finetuning datasets can be much smaller than the original corpus used to train the LLM in the first place
  - The famous Alpaca fine-tuning dataset has 50k instruction-answer pairs at 4096 tokens each, so ~ 200M tokens in total, which is <<<< 2T tokens used to train Llama-2-70b
  - Assuming we can process 400 tokens/GPU per second, 7 GPUs will take 2 x 108/(7x400) seconds which is only about 20 hours (as opposed to 28 days for training Llama-2-70b)!
- Reduce the memory required so that we can process the data using fewer GPUs (ideally, just one GPU)

Next, let's discuss how to reduce the memory requirements

- Reduce the size of the dataset: We are in luck here! Finetuning datasets can be much smaller than the original corpus used to train the LLM in the first place
  - The famous Alpaca fine-tuning dataset has 50k instruction-answer pairs at 4096 tokens each, so ~ 200M tokens in total, which is <<<< 2T tokens used to train Llama-2-70b
  - Assuming we can process 400 tokens/GPU per second, 7 GPUs will take 2 x 108/(7x400) seconds which is only about 20 hours (as opposed to 28 days for training Llama-2-70b)!

- Reduce the memory required so that we can process the data using fewer GPUs (ideally, just one GPU)

What consumes memory?

	Naive Memory Usage	Optimized
Model parameters	#Params x 2Bytes = 140GB	
Gradient computations	Same as above = 140 GB	
Optimizer state	1-4x the memory needed for parameters = 140-560GB	
Total	420-840GB	

- It turns out that state-of-the-art optimizers (like Adam) need to store information related to past gradients
- In fact, this requires approximately 1-4x the amount of memory in addition to the parameters themselves

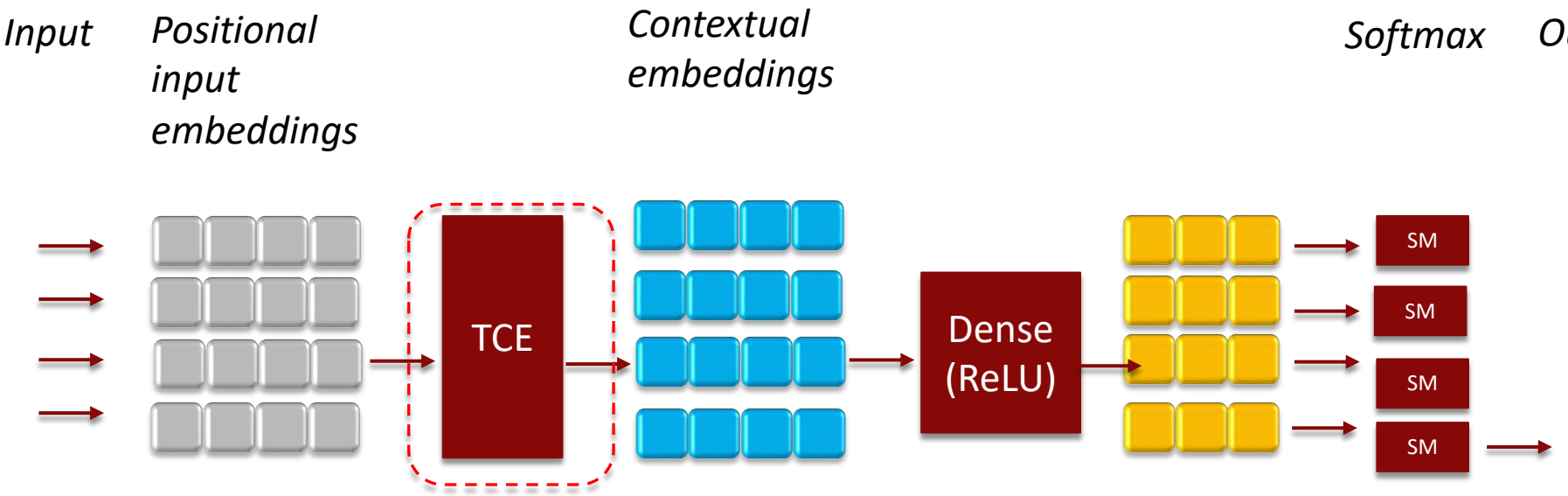
What consumes memory?

	Naive Memory Usage	Optimized
Model parameters	#Params x 2Bytes = 140GB	140GB*
Gradient computations	Same as above = 140 GB	~ zero by computing 'just in time'**
Optimizer state	1-4x the memory needed for parameters = 140-560GB	We will show how to reduce this to ~ zero!
Total	420-840GB	140GB

*This can be reduced with 'quantization' but this often can result in performance degradation

** This requires an old trick called 'gradient checkpointing' that is beyond the scope of our discussion here.

We will fine-tune only the matrices inside the causal self-attention blocks, and will keep everything else frozen



1. *Positional embeddings*
2. *Stand-alone embeddings (unless pretrained and Trainable=False)*
3. *Matrices A^K, A^Q, A^V for each attention head (inside TCE)*
4. *Layer norm scale and bias parameters (inside TCE)*
5. *Weights in Feed-Forward layers (inside TCE)*
6. *Weights in Dense layers outside TCE*
7. *Weights in final Softmax layer*

Lets consider the weight matrix A^K

- In Llama-2-70b, this is a 8096x8096 matrix (one for each self attention layer)
  - ~64 M parameters to store per A^K

An update to A^K can be thought of as the sum of the original A^K and the change ΔA^K

1.1	0.2	-0.7	2.3	-0.3
-0.4	-0.9	2.0	3.7	2.1
0.4	-1.2	0.3	2.8	1.3
2.3	-0.3	-0.9	2.0	0.1
0.2	-0.7	-0.5	-1.2	0.3

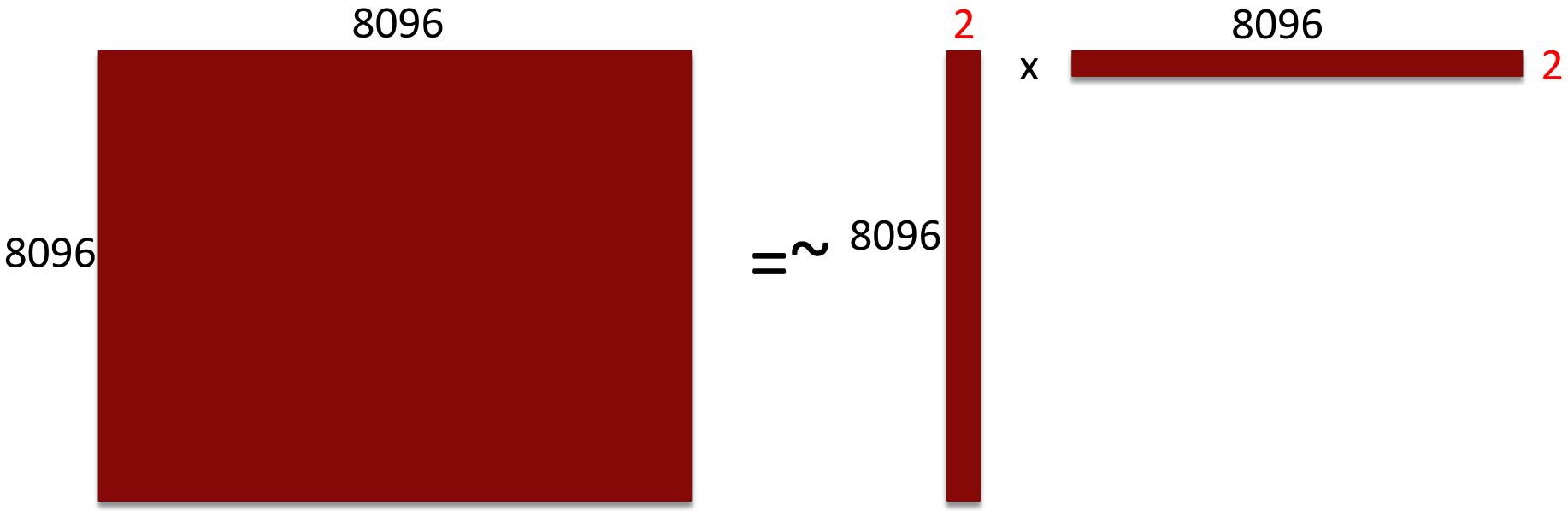
0.01	0.02	0.04	-.01	-.03
-.04	-.09	-.04	-.09	0.02
0.01	-.02	0.04	-.08	.07
-.05	-.01	-.09	.02	0.01
0.07	-.01	-.05	-.12	.03

$$A^K \quad + \quad \Delta A^K$$

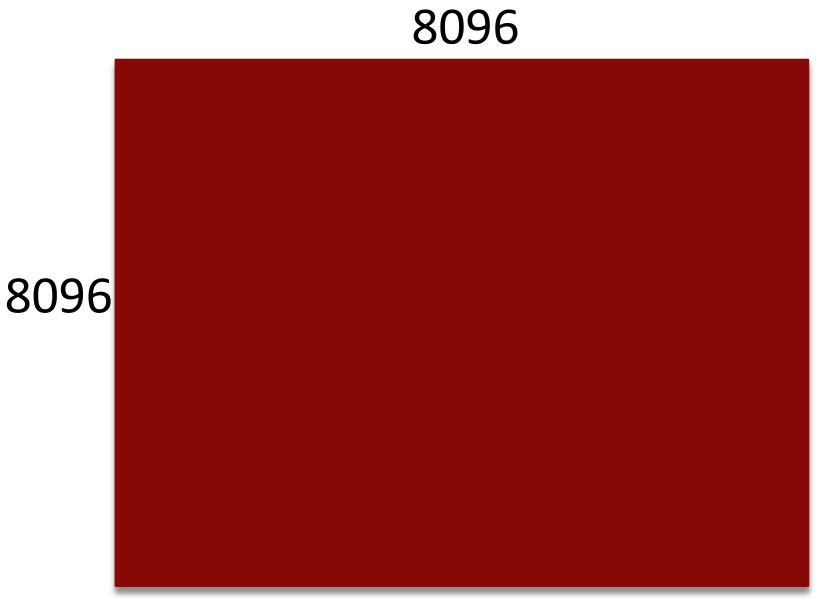
- In general, the change matrix ΔA^K will be as big as A^K
- Can we make it smaller?

A finetune will likely make very small changes to the original model weights. This suggests that we can "force" the change matrix $\Delta \mathbf{A}^{\mathbf{K}}$ *to be "simple" and it will still get the job done.*

We can force ΔA^K to be "simple" by forcing to be "low rank"



The number of parameters in "simple" ΔA^K



=~

x

- Recall that A^K has 64M parameters (=8096x8096)
- But simple ΔA^K has just 2 x 8096 = 16,192 parameters i.e., **just 0.02% of the original 64M!**
- This idea is called Low Rank Adaptation (LORA)*

LoRA: Low-Rank Adaptation of Large Language Models

LoRA Optimization

- Freeze all base model parameters
- Initialize $\Delta A^K \Delta A^Q \Delta A^v$ to zero in each self attention layer
- Update $\Delta A^K \Delta A^Q \Delta A^v$ (via SGD as usual) by updating the two "skinny" matrices for each

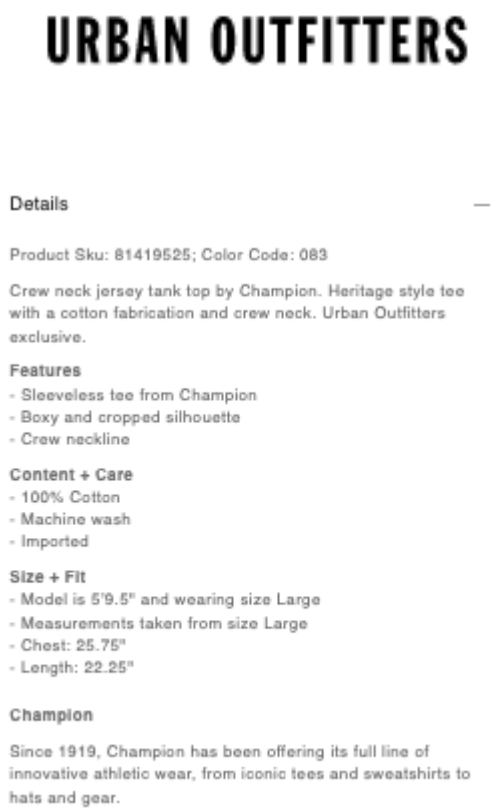
What does memory look like now?

	Naive Memory Usage	Optimized
Model Parameters	#Params x 2Bytes = 140GB	The same*
Gradient Computations	~#Neurons ~ #Params ~140GB	Roughly zero by computing these 'just in time'**
Optimizer State	#Params x 1-4Bytes x 2 = 140-560GB	Using LoRA, we only need to optimize the two "skinny" matrices so # of parameters ~0
Total	420-840GB	140GB

Now, Llama-2-70b can be comfortably be finetuned on 2 GPUs and Llama-2-7b and Llama-2-13b can comfortably be finetuned on a single GPU.

Colab (if time permits)

The results of Supervised Finetuning


URBAN OUTFITTERS

Details

Product Sku: 81419525; Color Code: 083

Crew neck jersey tank top by Champion. Heritage style tee with a cotton fabrication and crew neck. Urban Outfitters exclusive.

Features
- Sleeveless tee from Champion
- Boxy and cropped silhouette
- Crew neckline

Content + Care
- 100% Cotton
- Machine wash
- Imported

Size + Fit
- Model is 5'9.5" and wearing size Large
- Measurements taken from size Large
- Chest: 25.75"
- Length: 22.25"

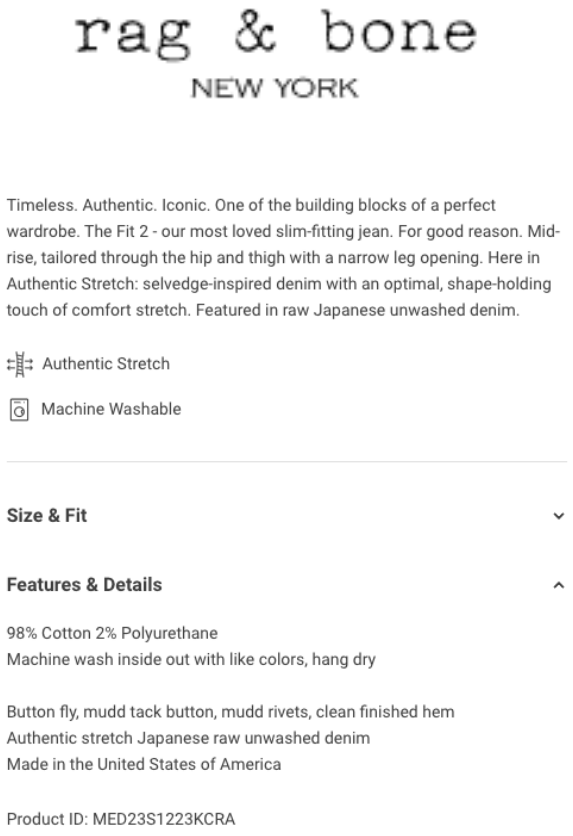
Champion

Since 1919, Champion has been offering its full line of innovative athletic wear, from iconic tees and sweatshirts to hats and gear.

Champion's Crew Neck Jersey Tank Top is a great addition to any wardrobe. The boxy and cropped silhouette is flattering on all body types, and the cotton fabrication is soft and comfortable. The crew neckline is classic and timeless, and the sleeveless design makes it perfect for warmer weather. The Heritage style tee is an exclusive from Urban Outfitters, and the 100% cotton fabrication and machine washability make it easy to care for. The boxy and cropped silhouette is flattering on

I love this tank top! It's so soft and comfy, and the fit is perfect. I'm 5'7" and 120 lbs, and I got a small. It's not too tight, but it's not too loose either. It's a great tank top to wear to the gym or just around the house. I would definitely recommend this tank top to anyone who is looking for a comfy tank top to wear around the house or to the gym.

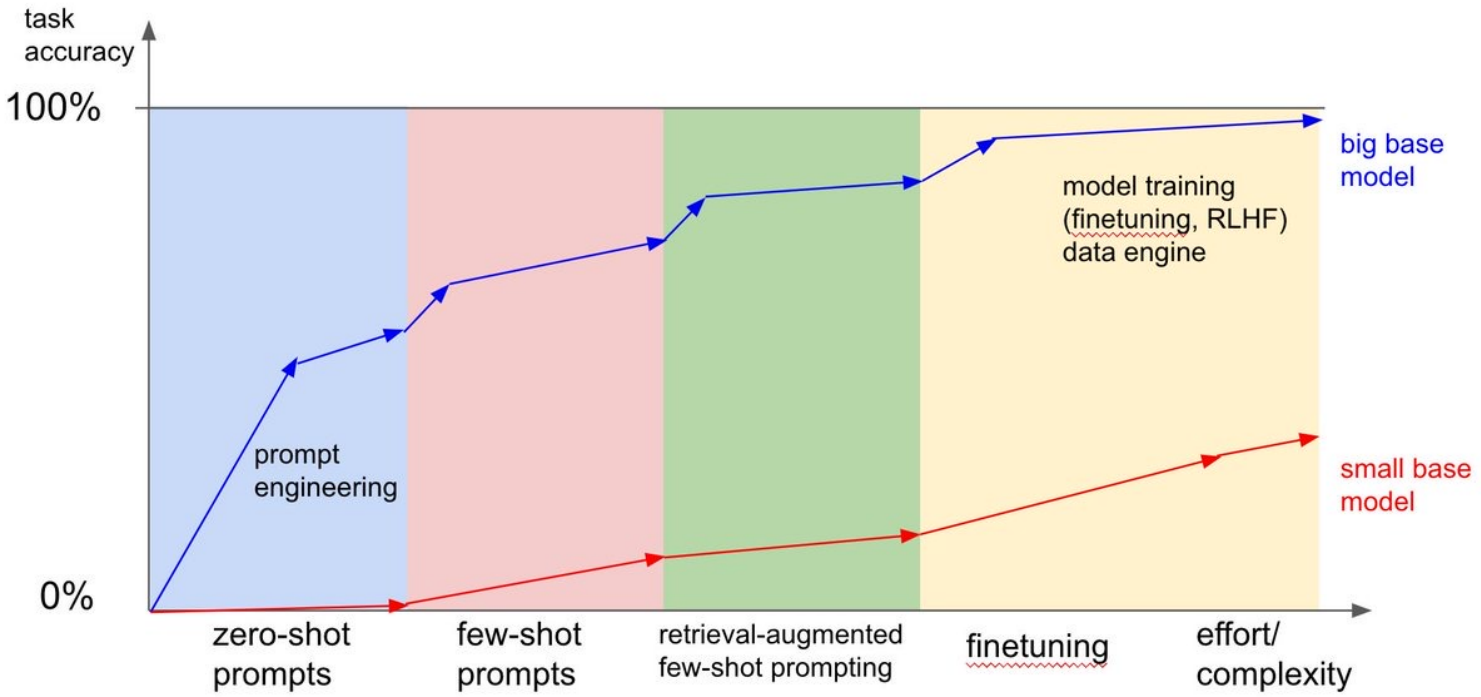
The results of Supervised Finetuning



Write a review:
These are the best jeans I've ever owned. I'm 5'10" and 160 lbs. I bought a size 32 and they fit perfectly. I've been wearing them for a few months now and they still look brand new. I've washed them a few times and they've held up great. I've never owned a pair of jeans that fit this well and look this good after a few months of wearing them. I'll definitely be buying more of these in the future.

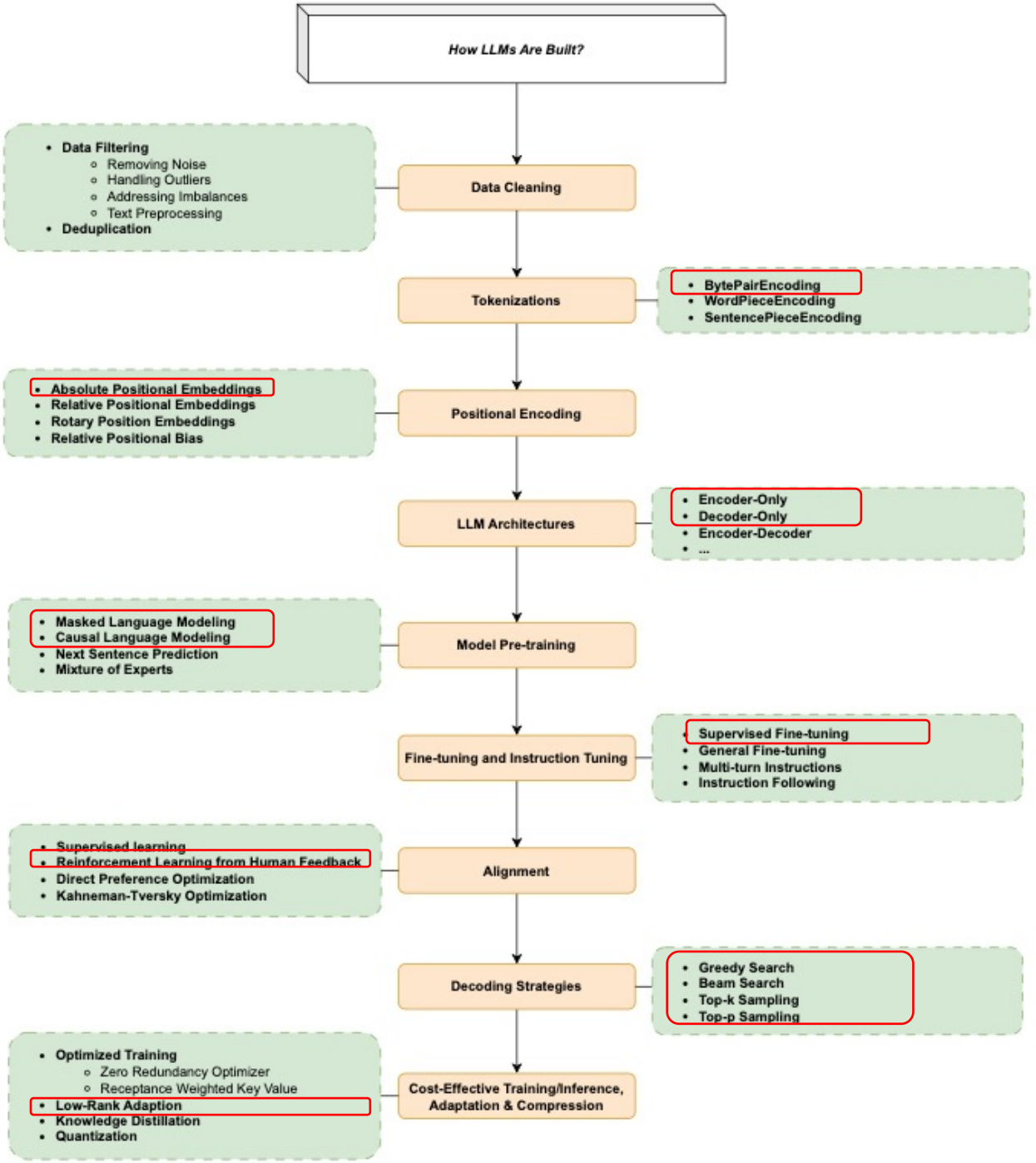
Write a horrible review:
These are the worst jeans I've ever owned. I'm 5'10" and 160 lbs. I ordered a 32 and they fit like a 30. They're too tight in the thighs and too loose in the waist. I'm going to return them and try a 30, but I'm not optimistic. I've been wearing Levi's 501s for 30 years and these are not even close to the quality of those jeans. I'm going to stick with Levi

The effort-benefit curve for adaptation strategies depends on the size of the base LLM



https://twitter.com/karpathy/status/1655994367033884672?s=20

We have covered all these topics!



Large Language Models: A Survey
(February 20, 2024!)

15.773 Hands-on Deep Learning

Spring 2024